Making Dictionaries

A guide to lexicography and the Multi-Dictionary Formatter

Software version 1.0

David F. Coward
Charles E. Grimes

SIL International
Waxhaw, North Carolina
2000

This book is sold with the software it describes. That software, too, is the copyrighted property of SIL International. However, in the interest of sharing the fruit of our research with the broader academic community, the user of the MULTI-DICTIONARY FORMATTER [MDF] software is granted the right to share copies of the distribution diskette with friends and associates, provided this is not done for commercial gain. Such recipients of the software, if they decide to use it in their research, should in turn buy this book with its latest version of the software.

MDF represents work in progress. In publishing this software, SIL International is making no commitment to maintain it. It is, however, committed to forwarding user comments to the software's authors, who may or may not develop the software further.

IBM is a registered trademark of International Business Machines Corporation. Microsoft Word, Microsoft Windows, Microsoft Word for Windows, and MS-DOS are trademarks of Microsoft Corporation.

Cover designed by Bud Speck.

The 2000 edition is only available in Portable Document Format (PDF). Only minor corrections to the 1995 text were made. No new material is introduced in this edition.

©1995, 2000 by SIL International
ALL RIGHTS RESERVED
Printed in the United States of America

ISBN 1–55671–011–9

Printed and distributed by:

JAARS, Inc.
International Computer Services (ICS)
Box 248, JAARS Road
Waxhaw, NC 28173
USA

Telephone: (704) 843-6085
FAX: (704) 843-6500

A catalog of publications of SIL International may be obtained from:

International Academic Bookstore
7500 W. Camp Wisdom Road
Dallas, TX 75236
USA

Contents

Preface vii

1. Before you begin 1
1.1 Installing the MDF program and files 1
1.1.1 Running MDF 1
1.1.2 Requirements and limitations 2
1.1.3 Further information 3
1.2 Notes on presentation and conventions 3
1.3 What to work on from the beginning 4

2. Getting started in lexicography with MDF 7
2.1 MDF fields used within an entry with the relative order in which they print 13
2.2 Examples of lexical entries (raw SHOEBOX form and MDF output) 29
2.3 Understanding the gloss, reversal and definition fields 36
2.3.1 Additional considerations for interlinearizing, definitions and reversal 41
2.3.2 Understanding the relationship between the \ge, \re and \de fields 43
2.4 Understanding the hierarchical structure of an entry 45
2.5 Direct character formatting within a field 49
2.6 Punctuation 52

3. Introduction to the Multi-Dictionary Formatter program 53
3.1 Familiarizing yourself with the program 53
3.2 Requirements and limitations 54
3.3 Overview of menu options 56
3.3.1 Change Settings 56
3.3.2 Reset 57
3.3.3 Format Dictionary 57
3.3.4 English and national language finderlists 60
3.3.5 Quit 62
3.4 Printing 63
3.5 Modifying the printout 64
3.5.1 WORD Stylesheets 64
3.5.2 Character Style codes 64
3.6 Summary 66

4. Basic strategies and perspectives 67
4.1 Terminology 67
4.2 Identifying the primary audience and purpose 68
4.3 Monolingual, bilingual, and trilingual dictionaries 70
4.4 Text-based lexicography and lexical sets of similar words 72
4.5 Minimal entries vs. expanded entries 74
4.6 Root-oriented vs. lexeme-oriented databases 77
4.6.1 Comparing the two approaches 83
4.6.2 Advantages and disadvantages 83
4.6.3 A suggested compromise 84

5. Structuring the database 89
5.1 Using a database structure vs. using unstructured text files in a word processor 89
5.2 Multiple language information (bilingual/multilingual lexical databases) 90
5.3 Categories of information in a lexical entry 92
5.3.1 Information about the headword 92
5.3.2 Information about words related to the headword 92
5.3.3 Housekeeping information 93
5.4 Sort sequences (alphabetizing) 93
5.4.1 Getting homonyms in the correct order 93
5.4.2 Restoring customized primary sort sequences 94
5.4.3 Sorting bound morphemes 95
5.4.4 Sorting citation forms (\lc) 96

6. Structuring information in lexical entries 99
6.1 Principles for choosing headwords 99
6.1.1 Affixes 103
6.1.2 Lexical root plus affixes 104
6.2 Choosing example sentences 105
6.3 Different words or different senses? (homonymy vs. polysemy) 107
6.4 Semantic categories (\sd, \th, \is) 115
6.5 Handling dialect information 117

7. Relating headwords to their lexical networks (lexical functions – \lf) 121

8. Considerations for special classes of entries 137
8.1 Folk taxonomies 138
8.1.1 Plants 142
8.1.2 Animals 144
8.1.3 Birds 146
8.1.4 Fish 147
8.1.5 Insects 147
8.1.6 Body part terms 148
8.1.7 Kin terms 148
8.1.8 Cultural items (artifacts) 150
8.1.9 Natural environment 151
8.2 Syntactic classes 151
8.2.1 Activities and events 152
8.2.2 States and processes 152
8.3 Loans and etymologies 153
8.4 Handling ritual speech and other special registers 154

9. Special considerations for parts of speech (\ps) 157
9.1 Common principles behind determining parts of speech 158
9.2 Common areas of discrepancy between principle and practice 159
9.3 Specific areas to watch out for 161
9.3.1 Views about the basis for assigning parts of speech 161
9.3.1.1 Are they adpositions or conjunctions? 161
9.3.1.2 Are they nouns or verbs? 162

9.3.1.3 Handling 'precategorials' (bound roots) 164
9.3.2 Verbal subclasses 166
9.3.2.1 Split-S (split intransitive) languages 166
9.3.2.2 Intradirective or quasi-reflexive verbs 167
9.3.2.3 Handling morphologically defined subclasses 168
9.3.2.4 Pragmatically motivated variants 169
9.3.3 Adjectives (versus nouns or verbs) 170
9.4 Summary of \ps issues 171
9.5 Checking paradigms (\pd) 171
9.6 Strategies for abbreviations 172
9.7 RANGE SETS (consistency check for sets of abbreviations) 175

10. Completing the dictionary 177
10.1 Extracting topical subsets (e.g. kin terms, plant terms) from the master lexicon for analysis or for separate publication 177
10.2 Writing an introduction to your dictionary 178
10.3 Acknowledgments for the dictionary 181

Appendix A: Alphabetized listing of field markers (with labels printed by MDF) 183

Appendix B: Relative order of fields in an entry (with labels printed by MDF) 187

Appendix C: Starter list of semantic domains (\sd) 191

Appendix D: Alphabetized starter list of lexical functions 193

Appendix E: Starter list of abbreviations 195

Appendix F: Enhancements and changes from v0.9 and v0.95 199
F.1 Enhancements in MDF v1.0 199
F.2 Changes from MDF v0.9 and 0.95 199
F.2.1 Changes in field markers 200
F.2.2 Changes in character formatting codes from v0.9x 206

Appendix G: Files and programs used by MDF 207
G.1 Print tables, etc. used by MDF 207
G.2 Programs required by MDF 208
G.3 Files created by MDF 208
G.4 Other files included on the release disk 208

Appendix H: Macros used in merging process 209
H.1 For WORD v5.0 209
H.2 For WORD v5.5 209
H.3 For WORD v6.0 210

Appendix I: Reporting problems or suggesting enhancements 211

Bibliography 213

Index 223

Preface

This book and the MDF program that accompanies it did not just grow in a vacuum. Rather the package developed as a positive response to a number of factors. It has been built on foundations laid by others. We acknowledge and thank them by reviewing the development process of MDF and this book (hereafter referred to as the *Guide*), noting their contributions where they happened.

David Coward worked closely with John Wimbish in the mid to late 1980s on the original development of the SHOEBOX computer program for data management. During the drafting of the initial SHOEBOX documentation Wimbish, Coward, and Grimes discussed the need to eventually rework and expand the chapter on lexicography and adapt it further as our experience and expertise grew. All three were working on genetically and geographically diverse languages in the province of Maluku in eastern Indonesia.

As the number of SHOEBOX users grew, many began to organize their lexical data and build dictionaries by interlinearizing bodies of vernacular texts. But it soon became apparent that there was a significant need for an easy way to format and print the dictionaries being compiled in SHOEBOX, and to produce a good reversed index. Coward developed a fairly complex CC (Consistent Changes) print table to print an early draft of his Selaru dictionary. Wyn Laidig and others then asked Coward to adapt similar tables for their needs—with many asking for refinements and enhancements to the original tables. It became obvious that one print table flexible enough to handle many options would be better than repeatedly customizing individual tables for individual users. Since many users of SHOEBOX were using their lexical database for both interlinearizing and building a dictionary, it also became apparent that there was a need for a conditional selection of information rather than a straight 'find-and-grab' approach for making a reversed finderlist (see §2.3). Because of the nature of the computer tools used for formatting and printing, these choices required superimposing certain constraints on the field codes within the lexical database, as undesirable as everyone knows that to be.

The development of the print tables was enhanced by the standards proposed and the issues addressed at the 1991 Hasanuddin University-SIL Lexicography Workshop in Sulawesi, Indonesia, lead by Tom Laskowske, Roger Hanna, Barbara Friberg, and Coward (as a guest). This included useful input from David Anderson and Phil Quick. The Maluku Linguistics Committee of SIL Indonesia, working at Pattimura University in Ambon, developed an enhanced set of suggested field codes. Bryan Hinton, Russ Loski, Howard Shelden, Mark Taber, and Ron Whisler were helpful at that stage, building on Wimbish (1989), the Sulawesi workshop, and the works of Len Newell (1986) and Marc Jacobson (1986). The results were made available in Indonesia in September 1992 as the

Maluku Dictionary Formatter [MDF] program (version 0.9, originally limited to feed into Microsoft WORD 5.0) with its accompanying documentation (Coward 1992). That version and the later v0.95 (for MS-WORD 5.5) quickly found eager testers in a number of countries throughout Southeast Asia and the Pacific. Many of these early testers provided helpful ideas and words of encouragement, and we especially thank Bryan Hinton, Jock Hughes, Rick Nivens, John Severn, and Ed Travis for theirs.

In the meantime, Grimes' responsibilities were taking him back and forth between Indonesia and Australia where he was gaining insights into semantics and related issues with Prof. Anna Wierzbicka, Prof. Bill Foley, and Prof. Bob Dixon, and assisting Prof. Andrew Pawley with workshops and courses on dictionary-making. MDF v0.9 was incorporated into a number of SHOEBOX courses taught by Grimes at the Australian National University while he was a Visitor in the Department of Linguistics at the Research School of Pacific Studies. The correspondence between Coward and Grimes, beginning at that time, grew into the collaborative effort you now hold in your hands.

The enhancements of both the program and the documentation since v0.9 have focused on 1) providing more interactive options for the user; 2) making the field codes more broadly applicable to users outside Indonesia (hence the original name was changed from Maluku Dictionary Formatter to Multi-Dictionary Formatter); 3) making the field codes more systematic and mnemonic; 4) providing additional categories and options requested by early users working in a wide range of linguistically and geographically diverse languages; 5) tying MDF into the broader academic world of lexicography; 6) addressing background and methodological issues that are beyond the immediate scope of the MDF computer program but which are faced by anyone seriously grappling with cataloging the lexicon of a language, and 7) including around 200 real-language examples showing how to organize such things as homonyms, citation forms, multiple senses, various kinds of cross-references, dialectal information, loan words, multiple-language glossing, and other categories of lexical information, illustrating both the form it should take in a SHOEBOX-like database and how MDF formats the information for printing. The idea is that if users can see what an example looks like, they are then more likely to be able to adapt it to their needs. Over time the documentation expanded to what it is now, fulfilling the long-term goal of providing a stand-alone field guide that users can have with them when doing their fieldwork. Also included is a bibliography directing users to where they can find issues discussed at greater length.

As with the development of the MDF computer program, this *Guide* has also benefited greatly from the works of others. General sources in lexicography such as Zgusta (1971) and Landau (1989) broadened our horizons. Bartholomew and Schoenhals (1983) was particularly useful on principles for choosing good example sentences. Newell (1986) provided a helpful summary for, among other things, determining multiple senses. A lexicography workshop held at Cenderawasih University in Irian Jaya in 1985, run by Prof. Joseph Grimes of Cornell University provided an introduction to the works of Igor

Mel'chuk and the usefulness of lexical functions. That introduction grew into Chapter 7, which has also appeared in modified form as C. Grimes (1994). Joseph Grimes has also given us considerable encouragement and has suggested many useful modifications to both the MDF program and the *Guide* toward their latter stages of development. Prof. Andrew Pawley at the Australian National University, who took C. Grimes under wing in various workshops and courses on dictionary making, graciously allowed us to adapt some of his materials for this volume, particularly in Chapter 8. Chapter 9 addresses a number of issues that users have asked about and was presented in an earlier form at the 1992 Asia International Lexicography Conference (C. Grimes 1992).

From these and many other sources, and from our experience working on dictionaries, both our own and helping dozens of others, we have gleaned and condensed much of the information found in this *Guide.* The ideas have been generalized, streamlined and formulated into a package we are confident will be useful to many in both its theoretical and practical applications.

Along the way, John Wimbish and Dan Davis have individually encouraged our efforts and we are grateful for their support. Wimbish also commented on parts of this *Guide*. A number of other people have also given useful feedback including Myron Bromley, Les Bruce, Barbara Dix Grimes, Len Newell, David Snyder, and Peter Wang. While the over-all feedback has been overwhelmingly positive, recognizing the practical service and guidance that MDF provides, not everyone has been in full accord with all of our recommended approaches because of practices peculiar to their region that we do not encourage here for principled reasons. The beauty of both MDF and this *Guide,* however, is that they are flexible enough to handle a wide range of options even beyond the various competing approaches and options explicitly discussed or recommended here—it is truly a Multi-Dictionary Formatter.

Doyle Peterson has given consistent administrative support for this project as it developed toward its later stages. Jim Albright and Betty Eastman provided helpful editorial suggestions. Our wives and families have graciously tolerated several late-night-to-early-morning sessions, simultaneously believing in the usefulness of the MDF project and hoping we would finish it soon.

David F Coward, M.A.
Charles E. Grimes, Ph.D.
Waxhaw, North Carolina

1. Before you begin

Welcome to the Multi-Dictionary Formatter [MDF]! The MDF computer program that accompanies this *Guide* is designed to make formatting and printing dictionaries, and making a reversed index relatively painless. This *Guide* assists you in both how to use the MDF program and how to set up your lexical information in a database (such as those compiled using SHOEBOX) for formatting and printing through MDF.

CAUTION: If your lexical database does not use the standard field codes recognized by MDF, do not use this program yet. First convert your lexical field codes to this standard (as explained in chapter 2 of this *Guide*).

1.1 Installing the MDF program and files

The SETUP program will guide you through installing MDF on your computer. A hard disk drive is highly recommended. At the DOS prompt type `a:setup`, then press ENTER. If you are installing MDF from a different drive use the appropriate designation (e.g. `b:setup`). Respond to the screen prompts using the default suggestions if you are uncertain. We recommend installing MDF in its own subdirectory as suggested by the SETUP program, e.g. C:\MDF. Consult the README file on the release disk for additional information.

1.1.1 Running MDF

The MDF program is set up to work with WORD v5.0, v5.5, or v6.0 and WINWORD (v2.0 or v6.0).[1] In order to run, MDF needs to know the *filename* of your lexical database. So, if the name of your lexical database is LEXICON.DB, you would type:

```
C:\MDF>mdf lexicon.db                [if database is in the default directory]
C:\MDF>mdf \sawai\lex\lexicon.db      [include path if database is elsewhere]
```

The MDF program will ask you to specify the version of WORD you are using. (Use the arrow keys and <ENTER> to select it). If you prefer to specify this from the command line, the following exemplifies how to do it:

---

[1]If the user specifies WINWORD as the word processor, MDF will format, split, and convert the database files to WORD documents, but makes no attempt to merge them (because MDF cannot access WINWORD directly). The user will need to then exit MDF and load each document file into WINWORD manually for merging and printing. For WINWORD, formatted dictionaries are named DICTN*.DOC; English reversed lists are ENGLS*.DOC; and national reversed lists are NATNL*.DOC. Some WINWORD 6.0 users will prefer to merge the DICTN*.DOC files together by using the Master Document View and buttons, and then later remove the section breaks introduced by that process.

```
C:\MDF>mdf lexicon.db v5          (for WORD v5.0)
C:\MDF>mdf lexicon.db v55         (for WORD v5.5)
C:\MDF>mdf lexicon.db v6          (for WORD v6.0)
C:\MDF>mdf lexicon.db win2        (for WINWORD v2.0)
C:\MDF>mdf lexicon.db win6        (for WINWORD v6.0)
```

The MDF program can have trouble merging documents in WORD v5.5 and WORD v6.0 simply because the glossary files used by those programs assume a default keyboard setup for each version of WORD. If the user has configured the keyboard in WORD to be different from the default configuration, MDF may malfunction at the point where WORD is called. So, *test MDF on a small section of your lexicon to see that all is working well before trying to process your whole lexicon*.[2] If MDF does not work properly, exit MDF, reconfigure WORD to its default settings, and try MDF again. A file named MDFSAMPL.DB is provided with MDF for testing that your system is working properly.

For Windows users: Drag the MDF.BAT file to a Program Manager group; edit its properties (ALT+ENTER); and add the name (and path) of your lexical database to the command line. Also be sure the Working Directory is the same as the directory in which you copied all of the MDF files.

1.1.2 Requirements and limitations

MDF is *not* a sophisticated program![3] It requires some user care. Allow plenty of room for MDF to work—approximately four times the size of your lexical database. Trying this program on a floppy drive would be unwise. The MDF program reserves the filenames DICT*.*, ENGL*.*, and NATN*.* for its own use. Do not use these names for your own files as they are likely to be deleted. MDF must be able to find the MS-DOS program SORT.EXE (SORT.EXE is supplied with MS-DOS and is usually found in the C:\DOS subdirectory). If it is unable to find SORT (i.e. if C:\DOS is not in the PATH command in the AUTOEXEC.BAT file), the MDF program will not be able to run properly. To test if MDF will be able to find SORT, type DIR | SORT at the DOS prompt:

```
C:\MDF>dir | sort          [note: | = vertical bar]
```

If this gives an *alphabetized* listing of the files on the default directory then all is okay (the line indicating the amount of free disk space is also sorted to the top). If the files are *not* sorted alphabetically, this means that the SORT program is not accessible. You will need either to specify a path that makes SORT accessible, or to copy SORT to a place

---

[2]Testing a small portion of your lexicon before trying the whole thing is important not only for testing the interaction of the programs, but also for ensuring that the structuring of your lexical information fits within the parameters set for working with MDF (see chapter 2).

[3]That is, computerwise, although what MDF can deliver to the user is very powerful.

where it can be found (like to the directory where MDF and its associated files are located).

MDF must also be able to find your word processor. MDF assumes your word processor subdirectory is specified in the PATH command of your AUTOEXEC.BAT file and that your word processor is named WORD.EXE. If you have more than one version of WORD installed and have renamed the files (e.g. WORD5.EXE and WORD6.EXE), make sure the version you want to use with MDF is named (or renamed) to WORD.EXE. Make sure that particular subdirectory is added to the PATH command in AUTOEXEC.BAT. To check this, from the MDF subdirectory type:

`C:\MDF>word<ENTER>` [check WORD-for-DOS]

`C:\MDF>win winword<ENTER>` [check WORD-for-WINDOWS]

If your word processor comes up, then the setup is okay.

1.1.3 Further information

More information, including the differences between MDF version 0.9x and version 1.0, is available in the "Overview" option in the MDF program and chapter 3 of this *Guide*. Or WORD can be used to view the MDF.DOC file directly.

1.2 Notes on presentation and conventions

This *Guide* is a marriage between a practical academic manual on lexicography and a computer software manual. Users who are not familiar with the range of conventions found in software manuals will find the following summary helpful.

UPPER CASE letters are used in this *Guide* to indicate computer program names and acronyms (e.g. SHOEBOX, MDF, WORD) and computer filenames (e.g. MDFDICT.CCT, SRT.EXE).

SMALL CAPS are used to indicate keys on a keyboard (e.g. <ENTER>) or program menu functions (e.g. SHOEBOX JUMP feature, RANGE SETS, DATABASE TEMPLATE).

`Monospace font` (i.e. fixed width Courier font) indicates information that appears on the computer screen or information that you type:

```
C:\MDF>mdf \shoebox\lexicon\lexicon.db<ENTER>
```

Keyboard conventions: Key names connected by a *plus* sign [+] indicate a combination of keys (e.g. ALT+F6 indicates press the F6 function key while holding down the ALT key). Key names separated by a *comma* [,] indicate a sequence of key strokes (e.g. ALT+F,V indicates press the F key while holding down the ALT key, then press the V key). Angle brackets indicate pressing the key named, for example <ENTER>.

Cross-references to more detailed discussion elsewhere in this *Guide* take two forms. A cross-reference to an entire chapter is simply ‘see chapter 7’. A cross-reference to a specific section uses the symbol [§] as in ‘discussed in §4.6’ (meaning chapter 4, section 6).

Throughout this *Guide* are found special boxes beginning with ‘CAUTION’, ‘TIP’, ‘NOTE’. They alert the user to information that will make the compiling, formatting, and printing of a dictionary more trouble-free and rewarding.

Many *examples* are given throughout this *Guide* to illustrate the accompanying discussion and show how MDF processes information. Most are real examples from dictionaries in progress. The few English examples that are found are simply meant to illustrate a basic idea of how to manage the data and are not meant to portray theoretical tightness in their definitions—that is not what they are illustrating.

On-line helps: On the MDF release disk is a file called LXFIELDS.DB, which is designed as an on-line help in SHOEBOX for organizing lexical information to format and print through MDF. One can ask this file, for example, what is the **\sc** field? what is it for? and how do I organize information in that field? One can also look at this file for information on recommended order of fields, punctuation appropriate to a particular field, etc.

Sample database: Another file provided on the MDF release disk is MDFSAMPL.DB. This provides a SHOEBOX file of a number of lexical entries in the Selaru language of Indonesia. Some of the entries are simple and some complex, but they illustrate a range of different possibilities. This file can be called up into SHOEBOX or a word processor and can be studied as desired. It can also be used to gain familiarity with MDF by processing MDFSAMPL.DB using the various menu options available in MDF to view the variety of output options provided for the user. This can be done by typing:

```
C:\MDF>mdf mdfsampl.db
```

1.3 What to work on from the beginning

The compiler of a dictionary should plan on doing at least the following things during the years it takes between starting and finishing the dictionary.

1) When first learning how MDF interacts with your data, *make a test file* of 50–200 entries, both simple *and* complex, making sure that every field and record in it is organized along the lines required for MDF.

   Format this test file through MDF with the various options likely to be needed for your various audiences and purposes.

Make a *reversed finderlist* through MDF as you will be doing with the final product.

Copy the appropriate MDF stylesheet for your printer to MDFDICT.STY and print your test file.

Inspect every detail of the printout. Adjust the way lexical data is organized in your LEXICON.DB, and make minor adjustments to the stylesheet to get the resulting printout you desire.

2) *Edit* or *enter* the rest of your lexicon to conform to what you have learned from step one above.

3) We recommend making a *back-up of your entire lexical database on diskette* after every significant work session, or every 50 entries. It is safest to cycle two or three separate back-up disks. This way, if the most recent session results in a corrupted file, and this corrupted file is saved to a back-up diskette, there is a back-up of a previous session still available prior to the corrupted file.

PREVIOUS SESSION (3)	PREVIOUS SESSION (2)	TODAY'S SESSION (1)	NEXT SESSION
			Diskette A
		Diskette B	
	Diskette C		
Diskette A			

4) For safekeeping we recommend *mailing a back-up copy on diskette* of your entire lexical database at least once a year to some location other than your normal workplace.

5) We recommend making a *hard copy printout of your full lexical database* at least once a year.

6) We recommend that you *process your database through MDF* after every 100–200 new or newly edited entries. A new printout is not required, just inspection of the results on the computer. This keeps you mindful of how the field codes interact with MDF. It also helps you pinpoint a snag if the program should hang for some reason.

Once the compilers are ready to print the 'final' product, *they should plan on at least two passes*:

1) The first pass is a *printout* of the entire database using the options they want for the final form. This includes *both the dictionary and the finderlist*.

These printouts should be *carefully inspected entry by entry* to see that everything is as desired. Human experience suggests that it won't be.

Make any corrections on the original lexical database, not on the MDF output (i.e. make changes in the LEXICON.DB file, not in the DICT.DOC file)!

2) *After you have written your introduction to the dictionary* (see §10.2), then make sure the lexical database is consistent with what has been said in the introductory material and reprocess the corrected database file through MDF. Repeat the steps above, if necessary.

3) Using WORD, *post-edit* anything that MDF cannot control directly in the final DICT.DOC file. For example, a) remove the '(dateprint)' from the footers; b) make sure the section dividers that begin a new letter are modified to reflect special characters and digraphs as appropriate; c) if the national language-vernacular diglot, or triglot option is chosen, replace labels to conform to what is appropriate for the country in which the national language is spoken. (The Indonesian labels to be replaced are listed in Appendices A and B); d) if the national language-vernacular diglot option is chosen, replace *Kamus* (meaning 'dictionary') in the footer with whatever is appropriate.

2. Getting started in lexicography with MDF

Dictionary-making (lexicography) is a multifaceted process. It includes at least the following aspects:

1) ***Understanding the language(s)*** structurally, functionally, semantically, and socio-culturally.

2) ***Structuring the information***, such as kinds of information in an entry, codes, ordering of information in an entry, etc.

3) ***Inputting the information*** (compiling the lexical database) normally over a period of years. This is best begun in the earliest stages of contact with a language and continued throughout—much is gained by doing it this way.

4) ***Checking and refining*** information in the lexical database.

5) ***Manipulating the data*** for analytic or other purposes, such as extracting semantic domains, doing reversals, etc.

6) ***Output***: deciding the format and making changes.

7) ***Printing***.

8) ***Marketing*** and distribution.

A tool like SHOEBOX can very nicely assist with aspects 2–6 above. The Multi-Dictionary Formatter [MDF] and this *Guide* are designed to be used in conjunction with SHOEBOX to beef up 2–7, especially points 2, 5 (reversals), 6, and 7.

Putting dictionary information in a database structure rather than in word processor text files has significant advantages in the compiling, checking and formatting stages.[1] SHOEBOX has brought these advantages to new heights in a 640K DOS environment with features such as:

1) Fast searches in large lexical databases.

2) Easy comparison of non-adjacent entries and copying information from one to the other with the JUMP feature.

3) User-defined sort orders (e.g. **n** followed by **ñ, e** followed by **é**), and the ability to handle digraphs (**ng, ch, ll, mb, nd**).

---

[1] See a more detailed discussion of these advantages in §5.1.

4) The ability to search across separate databases (e.g. comparing different dictionaries of the same language, lexicons of different languages, and different domains of the same language).

5) The ability to check for consistency against a master list using the SHOEBOX RANGE SETS (e.g. parts of speech, semantic domains). This provides a quality control in the compiling stage.

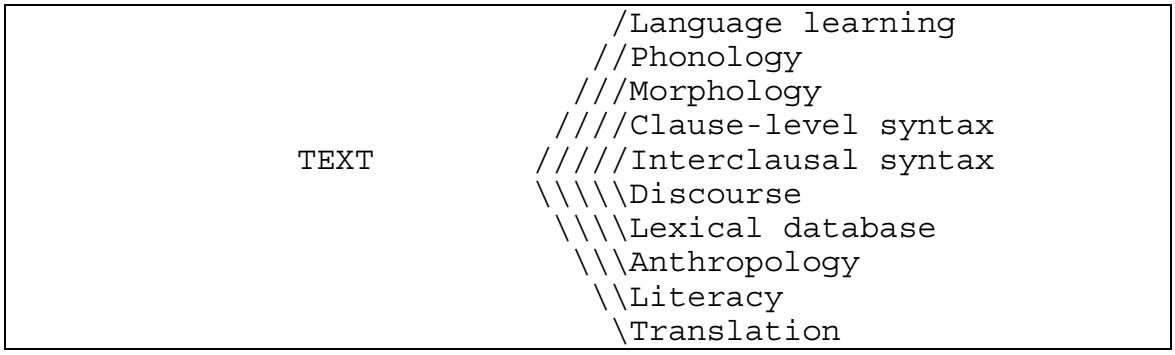
6) The use of a TEMPLATE for automatically inserting user-defined codes in a new entry.

7) The ability to manage housekeeping information as elaborately as needed without interfering with the printing or reversing of lexical information.

8) Storage of multiple language information and information for multiple purposes in the same place with one-time updating (e.g. glosses can be in the vernacular, English, national language, and regional language; and glosses can be designated separately for printing, for interlinearizing, or for reversing). This contrasts with updating the same material for different languages in separate files at different times, with the inconsistencies that result.

9) The use of SHOEBOX FILTERS to isolate or extract categories of information for analytical or special formatting purposes (e.g. part of speech, semantic domains, etymologies).

10) The lexical database is interactive with a text corpus (e.g. for interlinearizing, spell-checking, dictionary-building, or searching for example sentences). Text-based linguistics and lexicography provide a very sound foundation for mapping out a language and culture.



11) The ability to format semi-automatically, consistently and quickly. SHOEBOX allows user-defined codes.[2] Such codes can be systematically replaced by user-defined phrases, font, and style.

12) Database structures with a tool like SHOEBOX allow MDF to make a fairly sophisticated reversed finderlist in a short time, ranging from a few minutes to a couple of hours, instead of the weeks of busywork when done manually on word processor files.

The stages of formatting and printing a dictionary have been a continual source of frustration for many linguists and anthropologists who compile dictionaries using a database structure with standard format markers (backslash codes [\]) in a word processor or in SHOEBOX. Getting the information from a database format to a printed document can be so frustrating to the ordinary computer user that it may not get done at all—or at least not until one could get the help of a computer whiz. This difficulty is not limited to individual researchers compiling dictionaries semi-independently of technical support—the difficulty and frustrations are also shared by compilers of commercial dictionaries. For example, Landau (1989:29) observes that "dictionaries are notoriously difficult to typeset."

MDF is designed to bridge the gap between compiling and printing by enabling the average user to produce a double-column formatted dictionary from a standard format lexical database simply by pressing the letter **F** on the menu (for *Format dictionary*). By answering a few questions prompted by MDF, the resulting dictionary will have odd and even footers that include the name of the language and current date, section dividers with upper and lowercase letters between each new section of entries beginning with another letter, options of vernacular-English, vernacular-national language, triglot, and other outputs. By answering the screen prompts the user can get up to 16 different combinations without making any changes to the data file or to the MDF settings. Further combinations may be achieved by adjusting the MDF settings (through the CHANGE SETTINGS menu option and then following subsequent instructions) or the stylesheet (in WORD-for-DOS 5.0, 5.5, and 6.0). The compiler does not need to make any changes in their lexical database file, since MDF reads the information from the unchanged SHOEBOX LEXICON.DB file—ignoring SHOEBOX-internal fields and others (e.g. _no, \dt). The user thus does not need to remove these unwanted fields by other means.

Another menu option, **E** (for *English finderlist*), provides the user with a reversed finderlist that merges duplicate glosses and keeps track of which homophone and which sense the item refers to in the main dictionary. The primary menu options are as follows:

[2]With MDF the user will do best to stick with the suggested codes. Nearly 100 field codes are provided, covering most functional needs.

Multi-Dictionary Formatter
Overview **F**ormat dictionary **E**nglish finderlist **N**ational finderlist **C**hange settings **R**eset

Standard Format lexical database (e.g. SHOEBOX)

```
\lx dapan
\ps n
\ge spear
\de three-pronged spear  with
    barbs, used for eels
\ee This is similar to the
    unbarbed fv:nasel used
    for crayfish.[3]
\mr dapa-n
\dt 14/Apr/93
```

Formatted output [through MDF]

dapan *n.* three-pronged spear with barbs, used for eels. This is similar to the unbarbed **nasel** used for crayfish. *Morph:* **dapa-n**.

```
\lx flawan
\ps n
\sn 1
\ge gold
\et *bulaw-an
\eg gold
\dt 13/Dec/93
```

flawan *n.* 1) gold; 2) **majesty**. *Etym:* ***bulaw-an** 'gold'.

```
\lx akal
\ps n
\ge idea
\re idea ; notion ; conspiracy
\de idea, notion, conspiracy
\ee Has overtones of evil or
    mischievous intent.
\bw Arabic
\dt 20/Oct/89
```

akal *n.* idea, notion, conspiracy. Has overtones of evil or mischievous intent. *From:* Arabic.

A sample of MDF output for a formatted dictionary and a reversed finderlist are found on the following two pages:

[3]Note that in the \de field normal punctuation is used *except at the end*, where no punctuation is used—MDF will supply it later. The fv: is a code (font-vernacular) that provides direct formatting for printing the tagged word in the vernacular style when using MDF. Other direct formatting character codes are explained in §2.5.

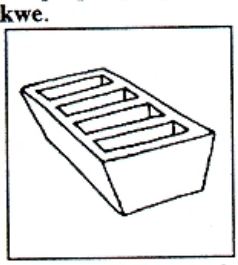
-ebat *v.* clean with a broom. *Ref:* n2.065.26 **Yebat bailola.** He cleaned the spiderwebs (with a broom). *See:* **-soly** 'sweep'; **-ury** 'rub'; **-kur** 'brush'. *Prdm:* 1. *1s:* **kebat**.

ebelurun *n:an.* pelican. *Ref:* d2.083.04 *See:* **manu** 'bird'. *Variant:* **ebelurungke, ebelurungare** (Younger speakers use 'ng'.). *Sg:* **ebelurunke**. *Pl:* **ebelurunare**.

ebnomatruma *n.* ancestors, descendant. *Ref:* d2.091.14 **Ebnomatrumakw rmatdye, ode lea kmwesan bo.** All of my ancestors have died, I am left all alone. *Ref:* d2.091.15 **Ara ebnomatruman.** Our ancestors. *See:* **ebu** 'master'; **matruma** 'ancestors'.

H - h

hatw *n.* 1) rock. *Ref:* d1.031.14 **Kal hatkwe ma katya askwe.** I grabbed a rock and threw it at the dog. *Sg:* **hatkwe**. *Pl:* **hature**. 2) magic artifact, talisman, charm. This is not restricted to just stones, can be gotten from all animals (kidney stones). Some people get pearls. *Syn:* **sok**. *1s:* **hatukkwe**.



3) baking form. *Ref:* d4.107.16 **Hatw raskyer ti ke.** The baking form. A fired, earthen baking form used to make **skyerker**. Often today, this is replaced by a cast iron griddle, much like a waffle iron, that is in the shape of four hearts.

-haw *vt.* 1) pound. *Ref:* d1.105.07 **Tawai ma kimakwet, desike bonyo tal ma thaw.** We dry it until it is almost dry, then we take it and pound it. *Ref:* d2.093.06 **Mal askwe ma mhwaw alyakwe ti neskwe.** Get the pestle so you can pound the rice in the mortar. *Syn:* **-tutu**. *See:* **asw** 'pestle'; **nesw** 'mortar'. *Prdm:* 3. *1s:* **khwaw**. 2) to stick, poke, prick. *Ref:* d2.113.13 **Mseak! Kali kbat desike kamnasanare ribun, mana khyaw o.** Look out! That lemon tree has a lot of thorns and they'll prick you. *Ref:* d4.043.09 **Knwaik nam mahaw lwawkkwe.** I pulled out whatever stuck in my foot. *Ref:* d4.109.02 **Maskye hyawa yaw.** The fish stuck me. This includes any prick whether from sharp fish spines or fins, or plant thorns. *See:* **kamnasan** 'thorn'; **hanar** 'thorn'.

haya *v.* row, paddle. *Ref:* dr.015.44 *Prdm:* 3?. — *n.* paddle. *See:* **-hesi** 'paddle'; **hesy$_2$** 'oar'. *Sg:* **hayake**.

-he tel *vi.* smart. *Lit:* 'to know language'. *Ref:* n1.099.11 **Sitke hye telke yobak klahke ma nya i.** The cat is clever at looking for mice to eat. *Ref:* d3.022.00 **Ibai kleti so mane rahetelke ma rabilak i ti a Tnebarne.** They went to the west and they became smart, smarter than us here in Tanimbar. *See:* **tel** 'language'; **hehe tun** 'etiquette'; **hehe tel** 'knowledge'. *Prdm:* 3. *1s:* **khwe tel**.

hesy$_1$ *n.* bridewealth, dowry. *Ref:* n1.111.08 **Mmwa ma tba ti siel waimw heskye.** Come on and we'll go pay your brother's bridewealth. This was explained as: "*Hal perkawinan bicara adat.*" *See:* **hesi** 'price'; **dolan** 'treasure'; **iliary** 'antiques'.

hesy$_2$ *n.* oar, paddle. *Ref:* n1.111.06 **Mal heskye ma khwesi aroke.** Bring the oar so I can paddle the boat. *Ref:* d4.109.16 **Hesy ma lahametkye.** A fast oar. *See:* **-hesi** 'paddle'; **haya** 'row'. *Sg:* **heskye**.

-het *vt.* chop. *Ref:* d2.031.02 **Ana mhwet laranke o mtubal.** If you chop the ground (with your machete, it) will become dull. *Ref:* d1.021.06 **Mal turike ma khwet akwe.** Give me the machete so I can chop wood. *Ref:* n3.027.21 **Simatare ode lwawtare itde khyeta ity.** They chop at one another with their arms and legs. This is a chopping or hacking motion usually done with a machete. It is not a very controlled cut. *Syn:* **-sin** 'split'; *Syn:* **-dew** 'chop on end'. *See:* **hetheta** 'beater bar (in weaving)'. *Variant:* **-hety** (Occassional form.). *Prdm:* 3. *1s:* **khwet**.

The summary of "cutting" motions for Selaru are:

-akrina: split in two lengthwise

crowd *n.* **geba edemen**;
v. **hiwi**.
crumble *vn.* **bono**.
crush *v.* **epsara apu**;
v. **esek$_2$**.
crustacean *n.* **doha**.
cry *n.* **enangit**.
cry out *n.* **hulun**;
vi. **emngaha**.
cure *vt.* **ba ruba geba**.
curl up *v.* **eflege**.
curse *v.* **eplagiik**;
vt. **eplumak**.
cuscus call *v.* **ik geke**.
cuscus nest *n.* **emhisi**;
n. **emhisin**.
customs *n.* **adat**.
cut *n.* **kawa**;
v. **fola**;
v. **foto$_1$**;
v. **hete**;
vt. **bisi**;
vt. **bobi$_1$**;
vt. **dasa**;
vt. **efasa**;
vt. **fasa** (1);
vt. **hange**.
cut field *vi.* **enogi**.
cut into sections *v.* **fola**.
cut off *vt.* **dafa**.
cut s.t. short *v.* **epture**.
cycle *n.* **fulan** (2).

D - d

damage *vn.* **ebregat**.
damaged *vn.* **ebrega**.
damar *n.* **enoi**;
n. **kisi$_2$**.
dance *n.* **cefal**;
n. **joget**;
v. **bailele**;
v. **jumak**;
vi. **kiki$_1$**.
dance (k.o.) *vi.* **cakalele**;
vi. **epkiki**.
dance type *n.* **enusi**.
dancing *vi.* **enusi-enusi**.
dangerous *vn.* **edakun**.
dare *vi.* **berani**.
daughter *n.* **anafina**;
n. **anamhuka**;
n. **anan** (1);
n. **anat**.
daughter-in-law *n.* **fin-emsawan**.
daughter's husband *n.* **emsawan**.
dawn *TIME.* **blola lea**.
day *CLASS.* **hari**;
n. **hanga**;
Ntime. **beton** (2);
TIME. **beto** (2);
TIME. **har$_2$**.
day and night *TIME.* **beto-lea**.
deaf *n.* **keben**;
vn. **kebe**.
dear *n.* **heren**;
vn. **doun**;
vn. **filin tirin**.
death *n.* **enmata**.
decayed *vn.* **bono**.
deceitful *n.* **abunawas**;
v. **kalfujik**;
vn. **falfujik**.
deceive *vt.* **abunawas**;
vt. **bodok**.
deceiver *n.* **geba goda**.
December *n.* **Desember**.
deception *n.* **eplalit**.
deceptive *vn.* **eprua**.
decide *v.* **esngoi-esngoik**;
vt. **esngoik**;
vt. **fasa** (2);
vt. **fasak**.
decide penalty *vt.* **fasa baut**.
decision *n.* **besluit**;
n. **bisluit**;
n. **efnasat**;
n. **esngoit**;
n. **keputusan**;
vi. **efnasa**.
declare *v.* **belwatak**;
v. **bletak**;
v. **enahu**;
v. **fala**;
v. **kaba**;
vi. **ebletak**;
vt. **kabak**;
vt. **kanauk**.
dedicate *v.* **aun**.
deed *n.* **empunat**.
deep *n.* **lalen** (4);
vn. **emkele**;
vn. **gore**.
deer *n.* **bijangan**.
deerfly *n.* **dango**.

2.1 MDF fields used within an entry with the relative order in which they print

Fields already factored into MDF are listed below. Sticking with these field markers will permit automated reverse indexing and printing. The relative order of the field markers is the one we recommend.[4] The following fields are critically ordered in relation to each other: **\lx \hm \lc \se \ps \pn \sn**. The *order* of the other fields is fixed in printing, but there is some flexibility for user preference in how the information can be organized on screen in SHOEBOX. For example, some users prefer **\sd** (semantic domain) near the front while others prefer it at the end.

> **CAUTION**: There is a potential cost in deviating from the canned package. MDF is not highly interactive, so do not expect to customize the output except in limited ways. Nevertheless, be assured that MDF provides a wide range of options that have proven capable of organizing diverse lexical information for a variety of purposes and from a variety of languages spoken in Asia, Africa, the Americas, and the Pacific.

The explanation of the field codes that follows is supplemented in §2.2 by examples from the Buru, Selaru, and Tetun languages of how these codes are used.[5] Subsequent chapters expand the discussion of many of these codes. A summary of the information below is available in a helps file supplied with MDF (LXFIELDS.DB) that can be on-line in SHOEBOX when needed.

\lx ***Lexeme***: also known as *lemma* or *headword* [**\lx tuat**]. This is the key field or record marker that SHOEBOX uses to keep one entry separate from another. Bound morphemes are listed with a preceding or following hyphen [**\lx -oli, \lx nara-**]. For some languages it may be acceptable to give an inflectable citation form, such as the H-form given in Tetun for inflectable verb roots [**\lx holi,** representing the paradigm **koli, moli, noli, holi, roli,** where the linguist would tend to identify the root **-oli** but the community thinks in terms of **holi**]. Multiple word or phrasal lexemes are common. Once SHOEBOX is set up in v1.2 or earlier, the user no longer sees **\lx**, but rather **Key:** at the top of the SHOEBOX screen [**Key: tuat**]. Version 2.0 uses the actual record marker field [**\lx tuat**]. See §6.1 for an expanded discussion on choosing headwords. This field is obligatory for each entry.

---

[4]The recommended order of fields is listed more succinctly in Appendix B. Different purposes and different audiences may require a different setup, but MDF is not designed to assist with customized output beyond the built-in options.

[5]See the SHOEBOX manual for alternate ideas on organizing lexical information. This current MDF *Guide* is designed to expand and enhance the discussion in the SHOEBOX manual relating to lexical databases and provides for a wider range of lexicographic needs.

CAUTION: *This* **\lx** *field must not be added within an entry/record.*

\hm ***Homonym/homophone/homograph***: [**\hm 1, \hm 2, \hm 3**]. Different homonyms must be in separate entries (see examples in §2.2). These will sort correctly and format as subscripts using MDF. See §6.3 for principles to distinguish between homonyms and multiple senses of a single lexeme. Use only if needed. Cross-references to one of these entries should include the number, e.g. **\cf asw2**. When the file is converted to WORD format for printing, MDF will subscript the homonym number, e.g. *See:* **asw**$_2$. Where they occur, MDF automatically references the homonym number in the reversed finderlists.

\lc ***Citation form (lexical citation)***: [**\lx nara-, \lc naran**]. This gives a complete surface form of bound roots that will be printed as the headword in the final printout. The **\lc** form always replaces the **\lx** form for the printed dictionary. MDF prompts users to choose whether or not they want entries that use **\lc** to *sort* under the **\lc** form for the printed dictionary. If the entry is not sorted by the **\lc** form, it will *sort* under the **\lx**, but the *printed headword* will be the **\lc** form (**\lx -angu, \lc (na)-angu** is printed between **\lx ane** and **\lx aok**; similarly **\lx -ao, \lc (beke)-ao** is printed between **\lx aok** and **\lx ape**). See §5.4.4 for detailed discussion. Use **\lc** only if the **\lx** form is inappropriate for the printed dictionary. MDF places the contents of the **\lx** field as follows: **\lx -hilu, \lc na-hilu** is printed as **na-hilu** (*from:* **-hilu**).

\ph ***Phonetic form (pronunciation)***: An indication of pronunciation is needed only where phonetic information is underdifferentiated by the practical orthography. MDF will supply square brackets and print the contents of the **\ph** field as monospace Courier font; [**\lx enaka, \ph e?naka**] is printed as [`e?naka`]. The information on how to interpret the phonetic pronunciation of the practical orthography should be explained in the introduction to the dictionary. SHOEBOX v2.0 can handle certain phonetic fonts on screen (see SHOEBOX manual). The **\ph** fields may also be used following the **\se** (subentry) field.

\se ***Subentry***: This field is used if one is organizing the lexicon primarily around the root morphemes rather than the surface forms. It is also used by some compilers for languages in which phrasal lexemes are common (e.g. *put out*) where the preference is not to list the phrasal lexemes as separate headwords. Phrasal lexemes can be organized as **\se** sections under the words that make them up. Polymorphemic forms or phrases are listed under **\se**, which is like the **\lx** field except that it occurs *within* the record (entry), marking the word (or phrase) as a form derived from or associated with the root. Following this field

would be all the fields that make up a typical lexical entry. There can be several **\se** subentries within a record (entry). Subentries can also have multiple senses within them. MDF begins each subentry at the beginning of a new line: [**\lx destroy, \se destroyer**]. For bilingual dictionaries of minority languages, many lexicographers prefer to not use **\se**, listing everything as main entries to make it easier for the naive user to find information. Upon reversal, both the **\se** form and the **\lx** form are referenced for a gloss listed under the **\se** form (e.g. **\lx sima, \ge hand, \se simake klarake, \ge palm** reverses on the subentry as 'palm —**simake klarake**, see: **sima**').

\ps ***Part of speech***: [**\ps vt, \ps n, \ps PREP, \ps PRO**]. This is used to classify the *vernacular* form, not the English or national language gloss. For example, the quality *fat* might be an adjective in English, but a verb in the vernacular language. **\ps** labels should be refined as one's understanding of the language grows. In other words, don't believe your early labels. Consistency in labeling is important. The RANGE SETS in SHOEBOX can help with this. There should be no final punctuation. MDF prints the **\ps** contents as italics (case is printed as entered in the original file) and adds a period [\ps vt → *vt.*]. See chapter 9 for a variety of relevant issues and Appendix E for a starter list of abbreviations. If more than one **\ps** is used in an entry (e.g. one sense as a noun and another as a verb), then MDF starts each new **\ps** within an entry or subentry at the beginning of a new line, dividing the entry into sections on the basis of the **\ps**. See §2.4 for how this fits into the structural hierarchy of an entry.

\pn ***Part of speech (national)***: [**\pn kkt, \pn kb, \ps ks**]. This is used to classify vernacular parts of speech, labeling them with terms common to national language dictionaries. Keep in mind that part of speech categories in the national language may not match part of speech categories in the vernacular (see chapter 9). Consistent labeling is important. Use SHOEBOX's RANGE SET feature for this field.

MDF requires that the **\pn** field follow the **\ps** field:

\ps n	(noun)
\pn kb	(the national abbreviation for 'noun')

CAUTION: If the order of these two fields is reversed, MDF will not format the dictionary output properly.

MDF will format the **\pn** field only if you specify that the output is for a national audience for either diglot or triglot formats. When a national audience is specified, the contents of the **\pn** field will replace the **\ps** field. But if there

is no **\pn** field or it is empty, the **\ps** field will be output for the national audience as well as for an English audience. This limits the need for redundancy for those labels that are the same in both languages. (See also **\ps** above.)

\sn ***Sense number***: This field is used to distinguish multiple sense of meaning, or minor senses [**\sn 1, \sn 2, \sn 3** → 1), 2), 3)]. Where an entry (or subentry) has more than one sense, this code gives the number and marks the beginning of each sense. There should be no closing parentheses or final punctuation in this field.

> **TIP:** Do not forget to also put **\sn 1** in records that have multiple senses.

Sense numbers can subdivide subentries (**\se**) and parts of speech (**\ps**). Each **\sn** should contain its own set of basic field markers (**\ge, \re, \de,** etc.) as relevant. It is important to aim toward each sense being validated by a well-chosen example sentence (**\xv**). See §6.2 and §6.3 for additional considerations. Where multiple senses occur, MDF automatically references the correct sense number in the reversed finderlists.

In compiling the lexicon, some lexicographers find it is convenient to deal with each separate language as a separate bundle (all English fields, then all national language fields), whereas others may prefer to interspersing the language codes (all the gloss fields, then all the reversal fields, then all the definition fields). See §2.3 for a discussion of the relationship between gloss, reversal, and definition fields.

Vernacular language bundle of fields:

\gv ***Gloss (vernacular)***: This field is primarily for a monolingual dictionary. It can be used as a temporary place to record succinct glosses provided by native speakers. For bilingual dictionaries the **\gv** information is best moved to the lexical functions fields (**\lf**) as Syn(onym), Ant(onym), Gen(eric), etc. (See chapter 7.)

\dv ***Definition/description (vernacular)***: Vernacular explanations or definitions of the headword generally should not be worded by the non-native speaker lexicographer. This field is for a monolingual dictionary and for retaining the integrity of native speaker explanations before they are repackaged in terms that make sense to the lexicographer.

English bundle of fields:

\ge ***Gloss (English)***: [**\ge 3s, \ge house ; hut ; building**]. This field is used for 1) interlinearizing, 2) printing the dictionary (if there is no **\de** field or the **\de** field is empty), and 3) reversal (if there is no **\re** field or the **\re** field is empty). Where the user is distinguishing morpheme-level from word-level glosses, the **\ge** field is used for *morpheme-level* glosses. Multiple word glosses should be connected with an underline to maintain spacing integrity and force SHOEBOX to treat the whole gloss as a unit when interlinearizing [**\ge put_out**, **\ge kin_group**]. MDF will convert this to a plain space when printing.

There are two options for organizing multiple glosses:

\ge house
\ge hut
\ge building OR

\ge house ; hut ; building [space-semicolon-space]

The SHOEBOX INTERLINEAR function can recognize either of these formats. For multiple glosses in either format MDF will separate them with comma-space. MDF also places a period after the final gloss. Thus, **\ge house ; hut ; building** is printed as: house, hut, building. The **\ge** field substitutes for a definition in printing a dictionary if no **\de** field is used. For speed in interlinearizing, the first gloss given should be the most common, broadest or most technical. It is not a definition! This field should be in all entries. See §2.3.

\re ***Reversal (English)***: [**\re jaw ; chin; \re exchange ; get ; take ; give**]. This gives the English word(s) or phrase(s) desired for a reversed English-vernacular finderlist. It is used for reversal only if the form in the **\ge** field is not suitable. The contents of the **\re** field are not printed in the dictionary, but only in the reversed finderlist. This is not a definition. Since this field is not used for interlinearizing, the joining underline [**\ge put_out**] is not used. See §2.3 for additional suggestions such as not glossing verbs as infinitives 'to (cut)', or nouns with an article 'a (rock)' because the reversal will sort on the first word in this field.

If an asterisk is placed in this field [**\re ***], then the relevant entry, subentry, or sense will be discarded or ignored for reversal (i.e. it will not be included in the reversed finderlist).

> **CAUTION**: MDF can handle *up to twenty* multiple glosses in the **\ge** or **\re** fields in a single sense or subentry for the reversal process. If more than twenty glosses are required, consider whether the information should be restructured into separate senses or subentries.

\we ***Word-level gloss (English)***: [**\we throw_out**]. If interlinearizing is desired at the word-level (surface form), rather than at the morpheme-level, then this field is used. See §4.6 for discussion of broader issues.

\de ***Definition/description (English)***: This field is used for a technical definition, expansion, or explanation of the meaning of the headword. It is more precise and complete than the gloss, aiming to capture *meaning* and aspects of *range* and *usage*. If there are **\de** field contents, then MDF will print them in the formatted dictionary and ignore the contents of the **\ge** field. In the **\de** field the compiler can reword or expand information in the **\ge** or **\re** fields using natural English worded for clarity for the broadest target audience. See §2.3 for examples and discussion of how the **\de** field relates to the **\ge** and **\re** fields. For additional overflow, use the encyclopedic fields (**\ee**) and usage fields (**\ue**). **NOTE:** Do not use final punctuation in this field. MDF will supply a period.

National language bundle of fields:

\gn ***Gloss (national language)***: This is like the English **\ge** field, but is for Indonesian, Spanish, French, Portuguese, etc. If interlinearizing is not to be done in the national language, then all material for a reversed finderlist is also put in this field and **\rn** is not used. See §4.2, §4.3 and §5.2.

\rn ***Reversal (national language)***: This is like the **\re** field, but is designed for forms that are appropriate for reversal in the national language. For example, **mempersilahkan** may be an appropriate gloss for the **\gn** field, but inappropriate for reversal—**\rn silahkan** is preferred. This field would also be used if interlinearizing is done in the national language and the contents of the **\gn** field are inappropriate for reversal.

\wn ***Word-level gloss (national language)***: This is like the **\we** field.

\dn ***Definition (national language)***: This is like **\de** field. If triglot printing is selected, national language fields are printed in italics.

Regional language bundle of fields: These are activated by MDF when National language audience or triglot options are selected.

\gr ***Gloss (regional language)***: This is like **\ge** field, but for the regional language or lingua franca that might be different from the national language, such as Ambonese Malay, Swahili, or regional creoles. These are often the languages in which explanations are given, particularly early in the researcher's contact, and they may provide more insight into the range of meaning of the headword than the national language. See §2.3, §4.2, and §4.3.

\rr ***Reversal (regional language)***: Like **\re** field. It is not likely to be needed.

\wr ***Word-level gloss (regional language)***: Like the **\we** field. It is not likely to be needed.

\dr ***Definition (regional language)***: This is like the **\de** field. If triglot printing is selected, MDF prints the regional language fields in italics within square brackets [ ] preceded by 'Regnl:' as in *[Regnl: parlente]*.

Fields clarifying the identity of the headword:

\lt ***Literally***: This is used where the literal parts of an idiom or lexeme do not obviously yield the gloss or definition given. MDF adds *Lit:* before the contents of this field and puts the contents in single quotes, followed by a period.

\sc ***Scientific name***: [**\sc Phalanger spp**]. Used where the information is known. Consult the best regional sources on flora, fauna, avifauna, and fish, or get expert advice. Be careful about guessing as a lay person. Educate yourself about principles of identification and taxonomy in botany and zoology. MDF prints the contents of this field as underlined italic, e.g. *Phalanger spp*. Do not use final punctuation as MDF will add this.

Example sentence bundle of fields: MDF can handle up to five different example sentence bundles for each sense and subentry in a main entry. Within such a unit, multiple examples are printed one after the other.

\rf ***Reference***: This refers to the source of the example sentences from data notebooks, the name of the source text and sentence number, etc. [**\rf C–89–2:34, \rf Manukama 164**.]. This housekeeping field does not have to be printed, but the information is useful to record. MDF adds *Ref:* before the contents of this field. The information is bundled with the following example sentence fields. Punctuation should be used as needed.

\xv ***Example (vernacular)***: Illustrative sentences in the vernacular legitimate and exemplify each separate sense. They should be short and natural. Examples extracted from texts may need to be adjusted to rebuild the information lost by removing them from their context. Punctuation and capitalization should be used as needed. Bartholomew and Schoenhals (1983: ch.9) have a helpful discussion of what makes good example sentences. See also §6.2. The contents of this field are printed in the vernacular font (i.e. bold).

\xe ***Example (English free translation)***: This is the English rendering of the example in **\xv**. Punctuation and capitalization should be used as needed. This field prints as regular font.

\xn ***Example (national language free translation)***: This is the national language rendering of the example in **\xv**. Punctuation and capitalization should be used as needed. In a diglot vernacular-national language dictionary the contents of this field print in italics.

\xr ***Example (regional language free translation)***: This is the regional language rendering of the example in **\xv**. Punctuation and capitalization should be used as needed. This prints only if the national language is requested.

\xg ***Example (gloss for interlinearizing)***: This field is for those who wish to include interlinear glossing of **\xv** in their lexicon.

> **CAUTION:** MDF does not currently recognize this field and so will not maintain the integrity of the spacing for printing if this field is used.[6] It is questionable whether interlinear examples are appropriate for most dictionaries.

Fields clarifying the range of meaning and usage:

\ue ***Usage (English)***: [**\ue archaic, \ue ritual, \ue Used by same-sex siblings, not opposite-sex siblings. \ue taboo, \ue vulgar, \ue Rana dialect, \ue H(igh register)**]. This is for comments on social usage, region, register, or dialect. It is also a place to note pragmatic connotations such as negative overtones if not clear from **\de** field. May overlap with lexical functions (**\lf**) such as SynT(aboo), SynD(ialect), or SynR(egister). Punctuation and capitalization should be used as needed. When printing, MDF places *Usage:* before the contents of this field.

\un ***Usage (national language)***: Like the **\ue** field.

\ur ***Usage (regional language)***: Like the **\ue** field.

\uv ***Usage (vernacular language)***: Like the **\ue** field.

\ee ***Encyclopedic information (English)***: This expands descriptive or ethnographic information in the **\de** field for outsiders who do not share the knowledge bank of the local community. The contents of this field are intended for printing (in contrast with the notes fields, such as **\nt**, which are not intended for final printing). Use normal punctuation and capitalization as needed.

---

[6]This reflects a limitation in the CTW program that MDF uses for converting to a WORD format.

TIP: Use the **\ee** and related fields (**\en, \er, \ev**) as all-purpose fields for anything that is not otherwise accommodated by the nearly 100 existing MDF field codes. MDF does not format the contents of the **\ee** field, but prints them as entered. MDF does not place an italic label before the contents of these fields.

\en ***Encyclopedic information (national language)***: Like the **\ee** field.

\er ***Encyclopedic information (regional language)***: Like the **\ee** field.

\ev ***Encyclopedic information (vernacular language)***: Like the **\ee** field.

\oe ***Only (restrictions—English)***: [**\oe human; \oe female; \oe not said for siblings of opposite sex; \oe collocates with non-active verbs only**]. This is for semantic or grammatical restrictions pertinent to the use of the headword. Capitalization should be used as needed. MDF places *Restrict:* before the contents of this field.

\on ***Only (restrictions—national language)***: Like the **\oe** field.

\or ***Only (restrictions—regional language)***: Like the **\oe** field.

\ov ***Only (restrictions—vernacular language)***: Like the **\oe** field.

Lexical function fields: This bundle of fields (**\lf \le \ln \lr**) should be kept together since each example of a lexical function has its own distinct glosses. There can be as many of these bundles as needed. MDF separates multiple bundles of lexical functions within an entry, subentry or sense with a semicolon [;], and places a period [.] after the final lexical function in the entry, subentry or sense.

\lf ***Lexical functions***: [**\lf Part = sufen, \lf Whole = huma**]. These are for mapping lexical networks, in effect, cross-referencing the lexeme with entries related to it, including various types of synonyms, antonyms, part-whole, generic-specific, typical actors, undergoers, instruments, material used, etc. The **\lf** system of cross-referencing links words in specific ways, in contrast to the use of **\cf**, where the link is vague and undefined. See the discussion of lexical functions in chapter 7 for a listing with examples of relations most commonly used in the **\lf** field. When printing, MDF converts the space–equals sign [=] to a colon [:], printing the label of the semantic relationship in italics, and what comes after the equals sign [=] as vernacular font. Thus, **\lf Syn = peni** prints through MDF as *Syn:* **peni**. MDF is set to ignore **\lf** fields that have nothing after the equals sign, for empty **\lf** fields that include certain labels in their

template. Thus, '**\lf Syn =** (blank)', will not print as *Syn:* unless something is filled in after the equals sign.

\le ***Lexical function (English gloss of \lf)***: [**\le merchant; \le wave**]. For most lexical functions, the contents of **\le** are simply the gloss of the contents of the **\lf** field. But for SynD(ialect), the dialect name is put in this field [**\le Rana dialect**]. For SynR(egister), the speech register name is put in this field [**\le Low**]. MDF places single quotes around the contents of this **\le** field. Thus, **\lf Nact** [Actor noun] **= gebkaleli, \le merchant** prints through MDF as *Nact:* **gebkaleli** 'merchant'. See §2.2 for examples of how these bundles are used.

\ln ***Lexical function (national language gloss of \lf)***: Like the **\le** field.

\lr ***Lexical function (regional language gloss of \lf)***: Like the **\le** field.

Additional fields relating the headword with its lexicocultural network:

\sy ***Synonyms***: Available for those who do not want to use the **\lf** bundles. This field does not provide the advantage of giving a gloss as with the **\le** field**.** MDF adds *Syn:* before the contents of this field and prints the contents in vernacular font, followed by a period.

\an ***Antonyms***: Available for those who do not want to use the **\lf** bundles. This field does not provide the advantage of giving a gloss as with the **\le** field**.** MDF adds *Ant:* before the contents of this field and prints the contents in vernacular font, followed by a period.

\mr ***Morphology***: [**\lx inaat, \mr ii-en-kaa-t**]. This field is for indicating morpheme representation, or the underlying forms where morphophonemic processes occur. MDF adds *Morph:* before the contents of this field and prints the contents in vernacular font, followed by a period. See §4.6 for further discussion with examples.

\cf ***Confer/cross-reference to other headwords:*** MDF converts this code to *See:* for the final printing, and the prints contents as vernacular font. Thus, **\cf anat** is printed as *See:* **anat**. This is a general purpose cross-reference that may, for example, be used in compounds to cross-reference the underlying roots [**\lx anrepun, \ge adopted_child, \cf repu**]. Complex instruments can be cross-referenced, e.g. *bow* with *arrow*, *mortar* with *pestle*, and vice versa. These can also be handled in the **\lf** field with the Counterpart [Cpart] relation. The **\cf** field is also used to cross-reference a minor variant to a main entry where fuller information is found (but see also **\mn** below). Cross-references to one of several homonyms should include the number (e.g. **\cf asw2**). When the file is

converted to WORD format for printing, MDF will subscript the homonym number (e.g. *See:* **asw**$_2$). MDF allows multiple **\cf** bundles, separating each with a semicolon [;] and placing a period after the final **\cf** bundle.

\ce ***Cross-reference (English gloss)***: Where the connection is not obvious it is helpful to have the gloss of the cross-reference in the entry at hand rather than have to chase it down [**\lx anrepun, \ge adopted_child, \cf repu, \ce retrieve**]. The contents of this field are printed in single quotes as in, *See:* **repu** 'retrieve'.

\cn ***Cross-reference (national language gloss)***: Like the **\ce** field.

\cr ***Cross-reference (regional language gloss)***: Like the **\ce** field.

\mn ***Main entry cross-reference***: This field is used to cross-reference a minor variant to a main entry where fuller information is found. It can also be used for a headword that reflects an unusual or irregular construction or inflection under which the user might look to refer to an entry where fuller information can be found. MDF adds *See main entry*: before the contents of this field and prints the contents in vernacular font, followed by a period [**\lx can't, \mn cannot**]. See **\va** below for a related field.

\va *Variant forms of headword*: [**\lx yako, \va ya, yak; \lx anat, \va an; \lx lidak, \va lidek; \lx cannot, \va can't**]. This can be the inverse of **\mn**. Cliticized forms, alternate pronunciations or alternate spellings are listed here. These variant forms generally refer to *minor entries* found elsewhere in the dictionary. Some lexicographers handle incomplete inflections or reduplication here as well, but those should be handled under the field(s) for paradigms (**\pd**) or reduplication (**\rd**). Use the **\ve, \vn,** and **\vr** fields only if there are relevant comments, such as distinguishing usage restrictions between the **\lx** form and the **\va** form. MDF adds *Variant:* before the contents of this field and prints the contents in vernacular font. Multiple **\va** field bundles are separated by a semicolon and the final bundle is closed with a period.

The **\va** bundle can also be used to record dialect variants.[7] See §6.5.

[7]We are aware that a compiler may use the **\va** bundle for more than one function (i.e. for morphological variants, and for dialectal variants), and that this sets up limitations for analysis or if one chooses to print one type but not the other. We intend future enhancements of MDF to have fields dedicated to dialectal information, but at present the programming limitations do not allow us any more field bundles. For the present, use **\va** and **\lf SynD =**.

\ve ***Variant (English comment)***: Comments regarding the contents of the **\va** field such as usage restrictions of the contents of **\va**, or dialect names identifying the source of the forms in **\va**. The contents of this field are enclosed in parentheses: **\lx hahy, \va fafy \ve older speakers**, prints as *Variant:* **fafy** (older speakers).

\vn ***Variant (national language comment)***: Like the **\ve** field.

\vr ***Variant (regional language comment)***: Like the **\ve** field.

Origins of the headword:

\bw ***Borrowed word (loan)***: [**\bw Sanskrit, \bw Swahili, \bw Spanish, \bw Malay**]. This identifies the ultimate source language, where known, with the understanding that it may have been introduced through an intermediate language. The form of the original language may also be given [**\lx emrimo, \bw Portuguese fi:meirinho**]. For the final printing MDF adds *From:* and places a period following the contents of the field, e.g. *From:* Sanskrit.

\et ***Etymology (historical)***: [**\et *biCuka, \et *maRuqanay**]. Reconstructed proto forms are given in this field. Cite attested published reconstructions only. Use **\nt** or **\ec** field if you want to posit your own guess at a reconstruction. MDF adds *Etym:* for the final printing.

\eg ***Etymology gloss (English)***: [**\eg bowels**]. This field is for the gloss of the reconstructed form so one can see semantic consistency or shift. Reconstructed meanings for most language families are given in English. Give the original published gloss—do not translate the published reconstructed gloss into the national language. MDF prints the contents of this field in single quotes, e.g. *Etym:* *biCuka 'bowels'.

\es ***Etymology source***: [**\es Blust 1993:46; \es PANDYMPL**]. This is for the source of the reconstructed form in **\et**. It is a housekeeping field for data management and is not intended for printing. Abbreviations for works on Austronesian languages can be found in Wurm and Wilson (1975).

\ec ***Etymology comment***: [**\ec metathesis, \ec Expect fv:lesun rather than fv:resun - possible loan**]. Relevant comments where the connection between the headword and the reconstructed form is not straightforward may be placed in this field. It may also be used to posit tentative unattested reconstructions and supporting data. Not intended for printing.

Grammatical paradigm fields:

\pd ***Paradigm***: This is a general field identifying the noun class, verb class, gender, or other paradigm set to which the headword belongs (as explained in the introduction to the dictionary). It can be used to identify incomplete or irregular paradigms. MDF places *Prdm*: before the contents of this field and adds a period at the end. For those users or languages that require more specific paradigm-related fields, MDF recognizes the following:

\sg	singular form	[*Sg:*]
\pl	plural form	[*Pl:*]
\rd	reduplication form(s)	[*Redup:*]
\1s	1st singular form	[*1s:*]
\2s	2nd singular form	[*2s:*]
\3s	3rd singular form	[*3s:*]
\4s	non-human or non-animate singular	[*3sn:*]
\1d	1st dual	[*1d:*]
\2d	2nd dual	[*2d:*]
\3d	3rd dual	[*3d:*]
\4d	non-human or non-animate dual	[*3dn:*]
\1p	1st plural	[*1p:*]
\1i	1st plural inclusive	[*1pi:*]
\1e	1st plural exclusive	[*1px:*]
\2p	2nd plural	[*2p:*]
\3p	3rd plural	[*3p:*]
\4p	non-human or non-animate plural	[*3pn:*]

Fixed format in field:

\tb ***Table (chart)***: This marks the text as unformatted. Line breaks and tabs entered by the user are retained. It may be used for such things as folk taxonomies of plants and animals, clarifying grammatical paradigms, or listing specific terms under a generic term (the latter better done in the **\lf** field). Punctuation and capitalization should be used as needed. The following example is from Selaru:

\tb Listing of all types of cutting verbs:
fv:akrina: split in two lengthwise
fv:boras: cut s.t. in small pieces with a knife
fv:dow: chop s.t. into smaller pieces while standing it on end
fv:het: chop or hack with a machete
fv:kety: slice open and clean an animal
fv:lary: slice (like chiles, etc.)

fv:lilit: shave or carve
fv:mair: to adze wood
fv:simat: pop out or cut out coconut meat

[MDF *prints this out as:*]
Listing of all types of cutting verbs:
akrina: split in two lengthwise
boras: cut s.t. in small pieces with a knife
dow: chop s.t. into smaller pieces while standing it on end
het: chop or hack with a machete
kety: slice open and clean an animal
lary: slice (like chiles, etc.)
lilit: shave or carve
mair: to adze wood
simat: pop out or cut out coconut meat

Alternatively these could be listed under a generic cutting verb in the **\lf** field as **\lf Spec = akrina, \le split in two lengthwise**, etc.

Tables may require some "tweaking" to fine-tune the formatting when the time comes to print the dictionary after MDF has ported the lexical file into MS-WORD.

Fields relating the headword to others of similar categories: These are helpful for analysis.

\sd ***Semantic domain***: [**\sd Nkin, \sd Nplant, \sd Vcut, \sd Vspeak**]. The use and placement of this field marker within the SHOEBOX database is up to the user. Some who use it regularly tend to put it near the front of the entry. Some users place **\sd** directly following **\ps**, using **\ps** to indicate strict subcategorization (e.g. **\ps vt**), and using **\sd** to indicate selectional restrictions (e.g. **\sd Vcarry**). Here one tries to catalog the semantic categories relevant to the language, being careful not to let the English force or mask the vernacular categories. The use of this field greatly assists specialized analysis or extracting topical subsets of the whole lexicon (e.g. publishing a special fascicle on plant terms). Several domains can be listed in the one field, if relevant, or one can use a separate **\sd** field for each sense. The contents of this field are not ordinarily printed, as it is primarily for analysis. But if one chooses to print the **\sd** fields, MDF places them toward the end of the entry, preceding the contents of the field with *SD:* and follows the contents with a period. See Appendix C for a suggested starter list of semantic domains and optional renderings.

\is ***Index of semantics***: Some MDF users have requested this field for correlating vernacular terms with Louw and Nida's (1988) Greek-English 93 semantic domain categories (many with additional subdomains). While useful for some purposes (like translation of Greek-based materials), the compiler is cautioned to remember that these categories are an etic checklist that may have no relation to emic categories in the vernacular. This field could also be used for the Human Relations Area Files [HRAF] categories from the *Outline of cultural materials* (Murdock, et. al. 1982). A third system that could be used is that of Hashimoto (1977) which provides an etic list of semantic domains that is more compact than HRAF and less language specific than Louw and Nida. Reversing on this field would yield semantically related entries grouped under the various Louw and Nida, HRAF, or Hashimoto semantic domains. MDF precedes the contents of this field with *Semantics:* and places of period following the contents of the field.

\th ***Thesaurus (vernacular)***: [**\th utan**]. This field is for the vernacular *generic term* under which the headword is emically categorized by the people themselves. For example, in Selaru, *masy* 'fish' has a broader semantic range than English *fish* because it also includes sea mammals and crustaceans. Similarly, the Buru generic term *manut*, whose Austronesian reconstructed form is glossed as 'bird', in Buru includes bats and other flying creatures like butterflies whose wings are large enough and slow enough to see in flight, but does not include most other insects. (See §8.1 for a discussion on folk taxonomies). This field is useful for later analysis or extraction (using SHOEBOX FILTERS) for separate publications of fish-type terms, flying creatures, etc. The contents of this field may or may not correlate with a western taxonomy or with the **\sd** field. It overlaps with **\lf Gen(eric) =**. MDF precedes the contents of this field with *Thes:* and places of period following the contents of the field.

Fields relating the entry to external material:

\bb ***Bibliographical reference***: [**\bb BDG 1991:328, \bb Schut 1917**]. This field references literature expanding on this lexeme. It is generally for grammatical particles or lexemes of ethnographic significance. MDF places *Read:* before the contents of this field and places period after.

\pc ***Picture***: This may refer to a sketch in a notebook, a photograph or slide in the lexicographer's collection, a picture or photograph in a published book, or a link to a computerized graphic file (e.g. file.PCX). If the field begins with *.G.*, then MDF will set it up in WORD to print as a graphics image in that entry.

```
\pc .G.\pcx\eagle.pcx;1.5";1";PCX
```

The .G. marks this as a graphics link. Next follows the path and filename: \pcx\eagle.pcx. Then the width of the picture desired for printing (here 1.5 inches), then the height (1”), and finally the graphics format type (PCX). Each bit of information is separate by a semicolon [;].

When the dictionary is formatted, the graphics information is moved to the beginning of the entry, subentry or sense in which the **\pc** field is found. This will cause the text to flow around the picture, which will be in a box. Sizes much larger than 1.5” x 1.5” are not recommended. In double column format the picture is placed flush right in the column; in single column format the picture is flush right to the right margin.

If no *.G.* is found, then MDF assumes the contents of the field are a reference to a book or notebook and simply prints the contents of the field enclosed in parentheses.

Note fields:

\nt ***Notes***: This is a general note field that can accommodate comments related to any field. It may be placed anywhere within an entry, subentry, or sense. Punctuation and capitalization should be used as needed. If selected to print, the contents of this field will be placed at the end of the entry or sense within square brackets [*Note: ...*]. These fields are intended for the compiler’s use and are not intended for printing, except for drafts. If the lexicographer wants to distinguish different classes of notes, MDF recognizes the following fields:

\np	notes—phonology and morphophonemics	[*Phon: ...*]
\ng	notes—grammar	[*Gram: ...*]
\nd	notes—discourse	[*Disc: ...*]
\na	notes—anthropology	[*Anthro: ...*]
\ns	notes—sociolinguistics	[*Socio: ...*]
\nq	questions for further investigation	[*Ques*: *...*]

Miscellaneous housekeeping fields:

\so ***Source of data or information:*** [**\so informant’s name/initials, \so researcher’s name/initials**, **\so village name/code**]. This is important where a range of sources or several researchers or a team of compilers are involved in producing a dictionary. Normally not printed. When selected for printing, MDF places *Source:* before the contents of this field and a period after.

\st ***Status for editing or printing:*** [**\st no print, \st done, \st check**]. This field can be used to later exclude entries that the informants have specifically requested not appear (e.g. in the national language dictionary they may fear

abuse if certain sexual terms in the vernacular are known by immigrants or officials from other ethnic groups). It can also be used to flag entries that are considered fully edited or that need further editing prior to final printing. Not normally printed. When selected for printing, MDF places *Status:* before the contents of this field and a period after.

\dt ***Date entry was last edited:*** This housekeeping matter can be automated with the SHOEBOX DATESTAMP feature. It is not normally printed.

\?? ***Unknown fields***: Fields entered by the user that are not recognized by MDF are placed within square brackets at the end of the entry and preceded by a double question mark *[?? ...]*. These can be toggled to print or not print through the **C**hange Settings menu option (where they are called the '(huh)' fields).

2.2 Examples of lexical entries (raw SHOEBOX form and MDF output)

Some compilers could organize their data quite well if they were simply given a few visual examples of how somebody else structures similar information and how MDF formats it. A variety of examples are given below with little commentary. These should be sufficient for many to go a long ways in compiling their lexical database and printing it through MDF. Additional examples are sprinkled throughout this *Guide* along with detailed discussion of relevant issues.

SHOEBOX lexical database	MDF formatted output

`\lx stife`
`\ps vt`
`\ge pour`
`\dt 2/Nov/89`
	[*simple entry*]
	stife *vt.* pour.

`\lx srapa`
`\ps vt`
`\ge slap`
`\de slap with open hand`
`\dt 27/Aug/91`
	srapa *vt*. slap with open hand.

`\lx -angu`
`\lc na-angu`
`\ps v`
`\ge interwoven`
`\dt 29/Apr/93`
	[*citation form*]
	na-angu (*from:* **-angu**) *v.* interwoven.

`\lx sau`
`\hm 1`
`\ps vt`
`\ge sew`
	[*homonyms*]
	sau$_1$ *vt.* sew.

```
\lx sau
\hm 2
\ps n
\ge anchor
\lf Whole = waga
\le boat
\dt 17/Jul/93
```

sau$_2$ *n.* anchor. *Whole:* **waga** ‘boat’.

```
\lx sau
\hm 3
\ps n
\ge fruit
\de8 succulent fruit (various),
    including breadfruit,
    rose apple, guava and
    cashew fruit
```

sau$_3$ *n.* succulent fruit (various), including breadfruit, rose apple, guava and cashew fruit.

```
\lx ati
\ps vt
\ge twirl
\re twirl ; pick up with tongs
\de twirl, pick up s.t. with
    tongs
\lf Nact = anafina
\le woman
\lf Nug = bia
\le starch paste
\lf NugSpec = bia polon
\le sago paste
\lf NugSpec = mangkau polon
\le cassava paste
\lf NugSpec = bia mangkau
\le cassava paste
\lf Ninstr = atit
\le sago paste twirler, tongs
\et *atip
\eg pinch off
```

ati *vt.* twirl, pick up s.t. with tongs. *Nact*: **anafina** ‘woman’; *Nug*: **bia** ‘starch paste’; *NugSpec*: **bia polon** ‘sago paste’; *NugSpec*: **mangkau polon** ‘cassava paste’; *NugSpec*: **bia mangkau** ‘cassava paste’; *Ninstr*: **atit** ‘sago paste twirler, tongs’. *Etym:* *atip ‘pinch off’.

```
\lx atit
\ps n
\ge tongs
\ge twirler_(for_sago_paste)
\cf ati
\ce twirl
\sd Ninstr
```

atit *n.* tongs, twirler (for sago paste). *See:* **ati** ‘twirl’.

[8]SHOEBOX can be made to give hanging indents on the screen by setting the margins (for both v1.2 and v2.0 under EDIT MARGINS) to Hanging Indent 5. Some find this gives a more orderly appearance. Hanging indents in SHOEBOX do not effect the formatting in MDF.

```
\lx gebhaa
\ps n
\sn 1
\ge husband
\lt big person.
\lf SynD = namorit
\le Rana dialect
\sn 2
\ge clan_head
\lf SynD = tean elen
\le Rana dialect
\mr geba-haa
\cf haa
\ce big, important, loud
```

[*multiple senses*]

gebhaa *n.* 1) husband. *Lit:* 'big person.' *SynD:* **namorit** 'Rana dialect'. 2) clan head. *SynD:* **tean elen** 'Rana dialect'. *Morph:* **geba-haa**. *See:* **haa** 'big, important, loud'.

```
\lx emata
\ps vt
\sn 1
\ge kill
\re kill ; murder
\rf C:89-3:27
\xv Siro rohi pa emata
    gebar telo dii.
\xe The two of them stalked and
    killed those three men.
\lf Nug = geba
\le person
\lf Nug = fafu
\le pig
\lf Spec = seka
\le spear s.o./s.t.
\sn 2
\ge extinguish
\rf B:86-1:84
\xv Da emata bana mele pothaki.
\xe She extinguished the fire
    lest it start a forest
    fire.
\lf Nug = bana
\le fire
\mr ep-mata
\cf mata
\ce die
```

emata *vt.* 1) kill. *Ref:* C:89-3:27. **Siro rohi pa emata gebar telo dii.** The two of them stalked and killed those three men. *Nug*: **geba** 'person'; *Nug*: **fafu** 'pig'; *Spec*: **seka** 'spear s.o./s.t.'. 2) extinguish. *Ref:* B:86-1:84 **Da emata bana mele pothaki.** She extinguished the fire lest it start a forest fire. *Nug*: **bana** 'fire'. *Morph:* **ep-mata**. *See:* **mata** 'die'.

```
\lx ahut
\ps n
\ge wave ; rough_(sea)
\ue Rana dialect.
\ee Rana speakers use fv:ahut
    to refer to rough seas when
    they are down at the coast,
    but it is taboo to use the
    term up at the lake.
\lf SynT = emhein
\le wave, rough (sea)
\lf Sim = permitek
\le stormy seas
\mr ahu-t
\et *qaRus
\eg current
\es PANDYPMPL
\sd Nnature
\dt 4/Mar/92
```

[*usage*]

ahut *n.* wave, rough (sea). *Usage:* Rana dialect. Rana speakers use **ahut** to refer to rough seas when they are down at the coast, but it is taboo to use the term up at the lake. *SynT*: **emhein** 'wave, rough (sea)'; *Sim*: **permitek** 'stormy seas'. *Morph:* **ahu-t**. *Etym:* *qaRus 'current'.

```
\lx fafu
\ps n
\ge pig
\re pig ; boar ; sow
\lf Spec = faf tinan
\le sow
\lf Spec = fafu bhasat
\le boar
\lf Spec = faf anan
\le piglet
\lf Spec = faf aba
\le wild (jungle) pig
\lf Spec = faf fena
\le domestic (village) pig
\lf Spec = fafu emlahat
\le domestic pig gone wild in
    the jungle
\lf Spec = fafu melaban
\le wild pig which has been
    domesticated
\lf Spec = faf Bali
\le short-legged domestic pig
    imported since WWII
\lf Spec = faf donit
\le fi:babirusa
\et *babuy
\eg pig
\es PAND
\sd Nanim
\dt 2/Nov/89
```

[*generic noun*]

fafu *n*. pig. *Spec:* **faf tinan** 'sow'; *Spec:* **fafu bhasat** 'boar'; *Spec:* **faf anan** 'piglet'; *Spec:* **faf aba** 'wild (jungle) pig'; *Spec:* **faf fena** 'domestic (village) pig'; *Spec:* **fafu emlahat** 'domestic pig gone wild in the jungle'; *Spec:* **fafu melaban** 'wild pig which has been domesticated'; *Spec:* **faf Bali** 'short-legged domestic pig imported since WWII'; *Spec:* **faf donit** '*babirusa*'. *Etym:* *babuy 'pig'.

```
\lx foto
\ps v
\ge take photograph
\ps n
\sn 1
\ge camera
\sn 2
\ge photograph
\bw English?
```

foto *v.* take photograph.
— *n.* 1) camera. 2) photograph. *From:* English?

```
\lx agat
\ps n
\ge grain
\de grain (generic)
\lf Nloc = hum kolon
\le grain bin
\lf Spec = feten
\le foxtail millet
\lf Spec = pala
\le rice
\lf Spec = biskutu
\le corn
\lf Spec = warahe
\le peanuts
\lf Spec = kopi [L]
\le coffee
\mr aga-t
\sd Nagri
\dt 2/Nov/89
```

[*multiple \lf bundles*]
agat *n.* grain (generic). *Nloc*: **hum kolon** ‘grain bin’; *Spec*: **feten** ‘foxtail millet’; *Spec*: **pala** ‘rice’; *Spec*: **biskutu** ‘corn’; *Spec*: **warahe** ‘peanuts’; *Spec*: **kopi [L]** ‘coffee’. *Morph:* **aga-t**.

```
\lx atet
\ps n
\ge thatch
\gn atap
\re roof ; thatch
\lf Sim = hum fafan
\le top of house, roof
\lf Mat = bia omon
\le sago leaves
\lf Mat = niwe omon
\le coconut palm leaves
\lf Mat = mehet
\le grass
\lf Prep = sau atet
\le sew thatch
\mr ate-t
\et *qatep
\eg thatch
\es PANDYPMPL
\sd Ncult ; Nhouse
\dt 23/Oct/89
```

atet *n.* thatch. *Sim*: **hum fafan** ‘top of house, roof’; *Mat*: **bia omon** ‘sago leaves’; *Mat*: **niwe omon** ‘coconut palm leaves’; *Mat*: **mehet** ‘grass’; *Prep*: **sau atet** ‘sew thatch’; *Morph:* **ate-t**. *Etym:* *qatep ‘thatch’.

```
\lx ama
\ps n
\pn kb
\ge F
\re father ; uncle
\de father, uncle; male of
    first ascending generation
    of ego's natal fv:noro or
    anyone ego's mother can
    call fv:naha 'brother'
\gn bapak ; ayah
\gr papi
\lf Gen = geba emtuat
\le parent, elder
\ln orang tua
\lf Spec = ama ebanat
\le birth father
\lf Spec = ama haat
\le father's oldest brother
\lr bapa tua
\lf Spec = ama roin
\le father's youngest brother
\lr bapa kacil
\lf Spec = ama kete
\le father-in-law
\ln bapak mertua
\lr bapa mantu
\lf Spec = ama tiri
\le stepfather (due to remarriage)
\lf Sim = tama
\le forefather of a lineage
\lf Cpart = ina
\le mother
\ln ibu
\et *ama
\eg father
\es PANDYPMPL
\sd Nkin
\dt 2/Apr/92
```

[*multiple language information*]

ama *n.* father, uncle; male of first ascending generation of ego's **noro** or anyone ego's mother can call **naha** 'brother'. *bapak, ayah.* [*Regnl*: *papi*]. *Gen*: **geba emtuat** 'parent, elder' '*orang tua*'; *Spec*: **ama ebanat** 'birth father'; *Spec*: **ama haat** 'father's oldest brother' '*bapa tua*'; *Spec*: **ama roin** 'father's youngest brother' '*bapa kacil*'; *Spec*: **ama kete** 'father-in-law' '*bapak mertua*' '*bapa mantu*'; *Spec*: **ama tiri** 'stepfather (due to remarriage)'; *Sim*: **tama** 'forefather of a lineage'; *Cpart*: **ina** 'mother' '*ibu*'. *Etym:* *ama 'father'.

[*Prints as above if triglot is selected for printing*. If diglot (English) is selected through the menu, then only the English and vernacular fields are printed. If diglot (National language) is selected, then the vernacular, national language, and regional language fields are printed, but the English fields are ignored.]

```
\lx kadefun
\ps n
\ge seat
\lf Syn = elepteat
\le seat
\lf SynL = kadera
\le chair, seat
\mr ka-defo-n
\cf defo
\ce stay, sit
```

kadefun *n*. seat. *Syn:* **elepteat** 'seat'; *SynL:* **kadera** 'chair, seat'. *Morph:* **ka-defo-n**. *See:* **defo** 'stay, sit'.

```
\lx ego
\ps vt
\ge transfer
\re transfer ; carry ; bring
    ; bear ; take ; get ; seize
    ; obtain ; grasp ; fetch
    ; marry
\de transfer control, location
    or affiliation; get, take,
    carry
\lf Spec = gao
\le grasp in hand, carry
    (e.g. a spear)
\lf Spec = wada
\le carry (bulky thing) on
    shoulder
\lf Spec = leba
\le carry on shoulder with a
    pole
\lf Spec = renge
\le carry s.t. in a basket on
    one's back using a tumpline
    (headstrap)
\lf Spec = eplabuk
\le carry on back using
    shoulder straps
\lf Spec = tolfafak
\le carry s.t. on head
\lf Spec = pinu
\le carry with strap over
    shoulder (e.g. hunting
    pouch)
\lf Spec = baba
\le carry a child on one's side
    with a carrying cloth
\lf Spec = sgera
\le carry a child with its legs
    straddling one's hip
\lf Spec = slolo
\le carry a child in one's arms
\lf Spec = sgege
\le carry s.t. under one's arm
\lf Spec = edaba
\le carry gifts on shoulder
    in procession
\sd Vcarry ; Vput ; Vexchange
\dt 28/Aug/91
```

[*multiple reversal units*]

ego *vt.* transfer control, location or affiliation; get, take, carry. *Spec*: **gao** 'grasp in hand, carry (e.g. a spear)'; *Spec:* **wada** 'carry (bulky thing) on shoulder'; *Spec:* **leba** 'carry on shoulder with a pole'; *Spec:* **renge** 'carry s.t. in a basket on one's back using a tumpline (headstrap)'; *Spec:* **eplabuk** 'carry on back using shoulder straps'; *Spec:* **tolfafak** 'carry s.t. on head'; *Spec:* **pinu** 'carry with strap over shoulder (e.g. hunting pouch)'; *Spec:* **baba** 'carry a child on one's side with a carrying cloth'; *Spec:* **sgera** 'carry a child with its legs straddling one's hip'; *Spec:* **slolo** 'carry a child in one's arms'; *Spec:* **sgege** 'carry s.t. under one's arm'; *Spec:* **edaba** 'carry gifts on shoulder in procession'.

```
\lx ba elalek
\ps vt
\ge faithful
\re faithful ; believe
    (strong sense)
\de faithful, believe
    (strong sense)
\lf Sim = nanuk
\le think, believe (weak sense)
\mr ek-lale-k
\cf lalen
\ce inside
```

[*cross-reference*]
ba elalek *vt.* faithful, believe (strong sense). *Sim*: **nanuk** 'think, believe (weak sense)'. *Morph:* **ek-lale-k**. *See:* **lalen** 'inside'.

2.3 Understanding the gloss, reversal and definition fields

A compiler can use a single lexical database for different purposes. For this reason it is useful to have several categories of gloss-type fields. We talk here about four: *gloss* fields (**\gv, \ge, \gn, \gr**), *reversal* fields (**\re, \rn, \rr**), *word-level* gloss (**\we, \wn, \wr**), and *definitions* (**\dv, \de, \dn, \dr**).

Gloss fields (\gv, \ge, \gn, \gr) and reversal fields (\re, \rn, \rr): It is important to understand that *glosses are not definitions!* Gloss fields are used for 1) interlinearizing, 2) making reversed finderlists (e.g. under what English or national language forms do you want to be able to look up this word?), if there are no reversal fields, and 3) getting a basic, but imprecise idea of the meaning of the word. This latter function is often called a 'translation equivalent', but would perhaps be better thought of as a translation approximation. Such glosses are often appropriate for use in translating the headword in some, but not all, contexts. Occasionally the same form can function for all these purposes and then only the **\ge** field is used.

```
\lx mitet
\ge black
```

mitet black.

```
\lx huma
\ge house
```

huma house.

However, there are several conditions in which the form of the gloss desired for interlinearizing is different from that desired for reversal. The first is when one form will suffice for all instances in interlinearizing, but several forms are desired for reversal, as illustrated below. Multiple options in the gloss fields slow down interlinearizing—a single form is inserted automatically, but multiple forms cause SHOEBOX to pause in order to let the user select the appropriate choice before resuming. Furthermore, in glossing strategies, a single form used consistently to gloss a word interlinearly (where legitimate) will more faithfully show the emic unity of a language than will a variety of

etic forms. Under these conditions the reversal fields are used to indicate the variety of surface forms desired for the reversal.

One emic form sufficient for interlinearizing	Dictionary	English finderlist

`\lx huma`
`\ge house`
`\re house ; hut ; building ; dwelling`
	huma house.	building **huma**
		dwelling **huma**
		house **huma**
		hut **huma**

`\lx aan`
`\ge jaw`
`\re jaw ; chin`
	aan jaw.	chin **aan**
		jaw **aan**

`\lx baa`
`\ge only`
`\re only ; exclusively ; just`
	baa only.	exclusively **baa**
		just **baa**
		only **baa**

A second condition is that in which an abbreviation is desired for interlinearizing, but not for the reversal. This is often desired for grammatical particles, but occasionally simply because the unabbreviated gloss would stretch out the interlinearization inordinately. In this case one would use the contents of the **\ge** field for interlinearizing, the **\re** for the English reversed finderlist, and the contents of the **\de** field for the printed dictionary.

Abbreviation preferred for interlinearizing:	Dictionary printout

`\lx saro`
`\ps PRO`
`\ge REC`
`\re reciprocal`
`\de reciprocal`
	saro *PRO.* reciprocal.

`\lx utan`
`\ps n`
`\ge veg.`
`\re vegetable`
`\de vegetable`
	utan *n.* vegetable.

A combination of both conditions of one form sufficient for interlinearizing and the preference for abbreviations for interlinearizing is common, particularly in certain semantic domains, or certain parts of speech:[9]

[9]Notice the preferred pattern for multiple glosses in either the gloss fields or the reversal fields of *space-semicolon-space* [gloss ; gloss]. This allows MDF to later convert these sequences to *comma-space* [gloss, gloss], without changing other sequences of *semicolon-space* [text; text] that are desired for other purposes.

Kin terms: Dictionary printout

```
\lx ina
\ps n
\ge M [mother]
\re mother ; aunt
\de mother, aunt; any female of
    the first ascending
    generation of ego's fv:noro
\lf Spec = infalin
\le mother's younger sister
\lf Cpart = ama
\le father
\sd Nkin
```

ina *n.* mother, aunt; any female of the first ascending generation of ego's **noro**. *Spec:* **infalin** 'mother's younger sister'; *Cpart:* **ama** 'father'.

Pronouns:

```
\lx ringe
\ps PRO
\ge 3s  [3rd pers. sing.]
\re he ; she ; it
\de he, she, it; third
    singular subject pronoun
```

ringe *PRO.* he, she, it; third singular subject pronoun.

Deictics or directionals:

```
\lx dii
\ps DEIC
\ge DIST  [distal]
\re that ; there ; then
\de that, there, then; distal
    in space, time, or
    reference
```

dii *DEIC.* that, there, then; distal in space, time, or reference.

Word-level glosses: These fields are used if the compiler needs morpheme-level glosses for some purposes and word-level glosses for other purposes. (See §4.6).

Definitions and the definition fields: *Definitions* represent a serious attempt to characterize the meaning of a lexeme in a precise way. Loose definitions tend to be expanded glosses or prose explanations of the lexeme. If present, the definition fields are printed in the dictionary. If not present, the contents of the gloss fields (**\ge**, etc.) are printed instead.

```
\lx ama
\ps n
\ge F
\re father ; uncle
\de father, uncle; male of
    first ascending generation
    in the fv:noro of ego's
    primary affiliation, or
    in the natal fv:noro of
    ego's mother
\ee Includes biological and
    classificatory 'fathers'
\sd Nkin
```

ama *n.* father, uncle; male of first ascending generation in the **noro** of ego's primary affiliation, or in the natal **noro** of ego's mother. Includes biological and classificatory 'fathers'.

True definitions, to a serious lexicographer, submit to certain theoretical constraints. Writing a good definition takes a lot of work. Consequently, in the process of compiling a lexicon it is common for far more lexemes to be just glossed, than to be both glossed and defined. Some scholars feel that formalism means precision, and so they tend to be algebraic in their 'definitions' (e.g. **leba**: x DO–*carry* y with **kalebat**, CAUSE y BECOME–at z). Occasionally such formalisms are motivated by desires to make a 'smart' dictionary for machine translation. However, such formalisms tend to be difficult for other dictionary users to understand or reproduce. They also tend to overlook other relevant information because they focus on only the kind of information encouraged by the particular formalism and discourage other kinds of information not accommodated by the formalism.

Many linguists and lexicographers distinguish between *denotative* meaning (a word's 'objective' referential meaning), and *connotative* meaning (the 'subjective' emotional associations with a word). Thus, (adapting Crystal 1985:88) in many western cultures *dog* has the denotative meaning of 'a canine quadruped', and its connotations including 'friend', 'companion', and 'helper'.

Traditionally lexicographers have tended to focus on denotative meaning at the expense of connotative meaning, partly because of prescriptive traditions about what constitutes scholarship and lexicography. However, a growing number of linguists and lexicographers are rejecting such a bipartite view of meaning, arguing that the *meaning* of a lexeme involves *both* denotative and connotative aspects. Hence to separate the two is artificial and academic, and definitions should include both aspects. These can both be included in statements in the **\de** field; or if the compiler feels uncomfortable blending the two, then connotative information can be encoded in the **\ee** (encyclopedic information) field, as in the entry for **ahut** in §2.2.

Some linguists use a 'natural semantic metalanguage' for definitions, limiting the words used in definitions to a set of 'semantic primitives and lexical universals' that form the building blocks for handling both connotative and denotative aspects of meaning as a

unified whole. Like any school of thought this requires an investment of time and energy to master and use well. The metalanguage may be awkward for the uninitiated, but extremely powerful to those who become familiar with its use. Those who have invested in mastering it, however, must take special care not to lose the broader audience.

An example of using a natural semantic metalanguage is in Wierzbicka's (1991:100–104) summary her discussion of the Javanese term *étok-étok* (defined in Horne 1974:178 as 'to pretend'):

> I don't want to say what I think/know
> I don't have to say this
> I can say something else

The following principles are generally subscribed to in relation to definitions, several being particularly relevant to monolingual dictionaries:

1) Only words accounted for elsewhere should be used in a definition (in monolingual dictionaries). This does not necessarily mean that all words used in a definition should be themselves defined, because of the problem in principle #4.

2) Definitions should not be circular. For example, *sugar* should not be defined in terms of *sweet*, and then *sweet* also defined in terms of *sugar*; or *pain* should not be defined in terms of *hurt*, and *hurt* also defined in terms of *pain*.

3) Semantically complex things, events, or concepts should be defined by terms that are semantically more simple than the headword.

4) Eventually some words are found to be indefinable. These are occasionally referred to as 'semantic primitives', and occasionally 'lexical universals'.[10]

5) The word being defined should not be used as part of the definition.

6) As much as possible, definitions should use familiar, high frequency words rather than use obscure or archaic words or technical jargon.

7) The most fundamental or essential parts of the definition should be expressed first (e.g. genus, species, and primary differentiae). Expansions can follow (using **\ee**).

8) The form of the definition should match the part of speech of the headword for major word classes. Nouns should be described by noun phrases, and verbs by

[10]The strong view of lexical universals holds that all languages have a lexical explication of the declared set of universals. The weak view holds only that the set of so-called universals has been demonstrated to provide convenient building blocks for definitions, and have lexical explication in *most* languages.

verbal predicates. A common mistake is to characterize all verbs as infinitives (e.g. 'to buy'), even for languages that have no infinitive forms.

In the course of 'doing lexicography' with skilled native speaker assistants, it is helpful to get a vernacular definition (**\dv**) or explanation early. This often yields valuable insights that are otherwise elusive for writing bilingual definitions. The more often people work on formulating good definitions, the better they become at it.

2.3.1 Additional considerations for interlinearizing, definitions and reversal

People who have been trained in classical Indo-European languages often precede their glosses with helper words.

```
\ge to_sail                [BAD EXAMPLE—not a model!]
\ge a_sail
\ge to_comb
\ge a_comb
```

This pattern has many disadvantages: 1) it lengthens the gloss for interlinearizing; 2) it adds redundant information to what is already in the **\ps** field (the **\ps** field will tell whether it is a noun or a verb—the gloss does not need to repeat this); 3) it will not reverse under 'sail' as desired, but under 'to', resulting in possibly hundreds of verbs clustering under 'to', and hundreds of nouns under 'a' in a reversed finderlist; and 4) it usually misrepresents the vernacular form in the **\lx** field, which is seldom an infinitive. A routine to strip out the 'to' before reversal would have to be sophisticated enough to leave any legitimate 'to' in the gloss, or in other fields.

TIP: Always remember that the reversal process will sort on the first word of each gloss unit in the **\ge** or **\re** fields.

Thus, the reversal fields provide for reversing on more forms or more specialized forms than those found in the gloss fields. For example, English reversals might include, **\re basin, wash; \re aunt, maternal; bamboo sp**(ecies). A morphologically complex national language such as Indonesian could, for *membersihkan,* be entered as **\rn bersih, mem–*–kan** so the reversal will be indexed by the root in the national language finderlist.

In any lexical database there are probably certain records that should be excluded from the reversed finderlists. For example, minor entries might be excluded from finderlists (because they are variant forms and contain little information anyway, so there may be no point in referencing such an entry). Or the entry might be a functor of some kind which really can't be given a gloss that could serve as a decent reference form in a finderlist (e.g. **3sPOS**).

TIP: For each entry, subentry, or sense that you want excluded from the reversed finderlists, place an asterisk (*) in the **\re** and **\rn** fields. Many bound morphemes or minor entries (variants) do not need to be reversed, if the information contained in them is redundant with fuller entries.

```
\lx -a
\ps GEN
\ge 3sPOS
\re *
\de his, hers, its
\gn -nya
\rn *
\mn -na
```

This example is a minor entry. The main entry is **-na** and is referenced in the **\mn** (main entry) field. The fields **\re** and **\rn** each contain an asterisk indicating that this record is not to be reversed (i.e. not to be included in either the English or national language finderlists.)

If more than one **\re** or **\rn** field is needed in a record section, it can be done in one of two ways: either in separate fields, or separated by a semicolon with a space on each side:

```
\lx nelnyely                    OR    \lx nelnyely
\ps n                                 \ps n
...                                   ...
\rn kebersihan                        \rn kebersihan ; bersih, ke-*-an
\rn bersih, ke-*-an
```

In either case the reversing print tables would create two entries in the national language finderlist for **nelnyely**.

NOTE: The national language reversing process is completely separate from the English reversing process. This means that **\rn** fields operate independently from **\re** fields. So just because the compiler chooses to use two **\rn** fields (as in the example above) does not mean there must be two **\re** fields, and vice versa. For English, a gloss like '*cleanliness*' would be adequate for glossing text, defining the lexeme in a dictionary, and reversing the English list. The record would look like this:

```
\lx nelnyely                          \lx nelnyely
\ps n                                 \ps n
\ge cleanliness           OR          \ge cleanliness
\re                                   \re
\de                                   \de
\gn kebersihan                        \gn kebersihan
\rn kebersihan                        \rn kebersihan ; bersih, ke-*-an
\rn bersih, ke-*-an                   \dn
\dn                                   ...
...
```

One important comment about this record: the gloss *kebersihan* occurs twice in the record, once in the **\gn** field and once in the **\rn** field. This is necessary because the user wants to reverse on both *kebersihan* and *bersih, ke–*–an*. Once an **\rn** field is detected as having data (i.e. **\rn bersih, ke–*–an**), the reversing program ignores all **\gn** fields in that section of that record. The reversing program will not take information from both **\gn** and **\rn** fields out of the same section of a record. So, once the user decides to reverse on *bersih, ke–*–an*, then *kebersihan* must also be added (since in this case both forms are felt to be needed in the finderlist).

This restriction on **\rn**, **\gn** fields also applies to **\re**, **\ge** fields.

2.3.2 Understanding the relationship between the \ge, \re and \de fields

Three points summarize earlier information:[11]

1) Only the contents of the **\ge** field are used for *interlinearizing*.

2) The **\ge** and **\de** fields are used for printing the *main dictionary*. **\re** is ignored for this purpose. If there are contents to a **\de** field, then that will be printed in the dictionary entry, and the contents of the **\ge** field will be ignored. Otherwise the contents of the **\ge** field will be printed.

3) The **\ge** and **\re** fields are used for the reversed English *finderlist*. **\de** is ignored for this purpose. If there are contents to an **\re** field, then that will generate entries in the reversed finderlist, and the contents of the **\ge** field will be ignored. Otherwise the contents of the **\ge** field will be used.

An important advantage to this conditionally sensitive or 'cascading' method of dealing with these sets of fields is that each lexical entry is not required to have all the fields filled in (**\ge, \re, \de, \gn, \rn, \dn**). This allows the fields to be used for the purposes needed without having excessive duplication where not needed.

To recapitulate the way MDF works, unless the settings are changed it will ignore the **\re** field when formatting the normal dictionary. In a regular dictionary one does not normally want to see the abbreviations found in certain **\ge** fields (such as 3s, REC, veg., etc.). When formatting the dictionary for printing, if MDF finds a **\de** field (that contains data), MDF will ignore the **\ge** field. Thus, if there is information in either the **\ge** or **\re** fields that one does want in a dictionary entry, that information should be reproduced in the **\de** field, but worded and formatted naturally. In the following examples some of the definitions (**\de**) are cursory or preliminary (e.g. **aan, alih**) and some are precise and complete (e.g. **a, alik**).

---

[11]The same relationship described in these points here holds for the national language bundle of fields. English is isolated here for presentational clarity.

If one form will work for all three field functions (interlinearizing, dictionary, reversal), then only the **\ge** field should be used:

```
\lx aken
\ge gallbladder
```

aken gallbladder.

If the information in the **\re** field is desired in the main dictionary, then it should be reproduced and reformatted in the **\de** field:[12]

```
\lx aan
\ge jaw
\re jaw ; chin
\de jaw, chin
```

aan jaw, chin.

If the information in the **\re** field is desired in a different form for naturalness, then the changes should be in the **\de** field.

```
\lx alih
\ge charge
\re charge (take)
\de take charge
```

alih take charge.

```
\lx bolo
\ge (bamboo)[13]
\re bamboo sp.
\de k.o. bamboo
```

bolo k.o. bamboo.

If more information is desired than is appropriate for the **\ge** and **\re** fields, then that should be in the **\de** field:

```
\lx ahut
\ge wave
\re wave ; rough
\de wave; rough (sea)
```

ahut wave; rough (sea).

```
\lx a
\ps PRO
\ge 1s
\re I
\de I; first person singular
    subject proclitic
```

a *PRO.* I; first person singular subject proclitic.

[12]The reason for choosing to not put both *jaw* and *chin* in the **\ge** field in this example is so that the SHOEBOX interlinearizing function can automatically fill in the gloss and move on. This is faster than having the program stop to ask the user to choose between *jaw* and *chin* each time **aan** is encountered in a text. If the stop-and-choose method is not seen as an inconvenience, then it is simpler to put both glosses in the **\ge** field and dispense with the **\de** and **\re** fields.

[13]Some find it convenient for interlinearizing to enclose a generic term in parentheses to indicate 'kind of x', thus avoiding multiple word glosses. Similarly *(name)* can be used as the gloss for a person's name, *(place)* for a place name, etc.

```
\lx alik
\ge peel
\re peel ; strip off (skin)
\de peel s.t. by hand with
    intent to use resulting
    core; strip skin or husk
    off s.t. by hand
```

alik peel s.t. by hand with intent to use resulting core; strip skin or husk off s.t. by hand.

It should now be clear that what one puts in the **\de** field is not limited to *definitions* in the strict denotative sense.

2.4 Understanding the hierarchical structure of an entry

Because of the nature of the computer tools that drive MDF, it has been necessary for MDF to superimpose a hierarchical structure that is flexible enough to meet most needs. The field codes that are relevant here are **\lx, \ps (\pn), \sn, \se**. Each of these sections or subsections can take a full set of field markers.

Multiple parts of speech (**\ps**) in an entry are used to organize sections within an entry. In many cases there is a clear relationship between a word functioning in different syntactic slots within a sentence as a noun, a verb, or a preposition, as between *shower* (v) and *shower* (n), and between *rain* (v) and *rain* (n). These are often clearly related to each other in meaning and have functional complementary distribution, and thus should not be handled as homonyms (see chapter 9 for a more detailed discussion of this and related issues). MDF starts a new **\ps** within an entry on a new line, preceded by an em-dash. If an entry is substructured in this manner, then sense numbers (**\sn**) are not needed unless to further substructure the part of speech (as in the second example below).

```
\lx anchor
\ps n
\de instrument attached to a
    rope or chain for
    preventing or minimizing
    the movement of a boat when
    it is not tied at dock,
    usually by friction along
    the ocean or lake bottom
\ps vt
\de action of using such an
    instrument
```

anchor *n.* instrument attached to a rope or chain for preventing or minimizing the movement of a boat when it is not at dock, usually by friction along the ocean or lake bottom.
— *vt.* action of using such an instrument.

Sense numbers (**\sn**) are also used to organize sections within an entry. Multiple senses should be grouped under the relevant parts of speech. Multiple senses in each separate part of speech should start with ‘1’.

\lx lexeme
\ps n \sn 1 \ge gloss \de definition \sn 2 \ge gloss \sn 3 \ge gloss \de definition

lexeme *n.* 1) definition. 2) gloss. 3) definition.

\lx lexeme
\ps n \sn 1 \ge gloss \de definition \sn 2 \ge gloss \de definition \sn 3 \ge gloss \ps v \sn 1 \ge gloss \sn 2 \ge gloss \de definition \sn 3 \ge gloss \de definition

lexeme *n.* 1) definition. 2) definition. 3) gloss.
— *v.* 1) gloss. 2) definition. 3) definition.

Some lexicographers want to make fine distinctions between *subsenses*. The principles for justifying subsenses are the same as those for justifying senses (see §6.3); the difference is one of degree or scope. Subsenses are more related to each other than they are to other senses. These can be handled in MDF in the **\sn** field with subcategorization using a, b, c, etc.

```
\lx opon
\ps n
\sn 1a
\ge grand_kin
\de grandparent, grandchild;
    reciprocal term of plus or
    minus two generations
\sn 1b
\ge ancestor
\de ancestor, descendant
\sn 2
\ge master
\de master, lord, owner; the
    one with the say over s.o.
    or s.t
```

opon *n.* 1a) grandparent, grandchild; reciprocal term of plus or minus two generations. 1b) ancestor, descendant. 2) master, lord, owner; the one with the say over s.o. or s.t.

Subentries (**\se**) provide a further level of hierarchy. These are commonly built around polymorphemic forms in a root-based dictionary (see §4.6 for extended discussion). Note that while information might be organized as follows during the early years of contact with a language, the information for 'brushcutter' below should eventually be separated out and placed elsewhere as it is not lexically related to this headword.

```
\lx brush
\ps n
\ge gloss
\de definition
\se hairbrush
\ps n
\ge gloss
\se paintbrush
\ps n
\ge gloss
\de definition
\se brushcutter
\ps v
\ge gloss
\de definition
\ps n
\ge gloss
\de definition
```

brush *n.* definition.
- **hairbrush** *n.* gloss.
- **paintbrush** *n.* definition.
- **brushcutter** v. definition.
  — *n.* definition.

```
\lx bersih
\ps adj
\sn 1
\ge clean
\de be clean, not dirty or
    messy
\sn 2
\ge innocent
\de be innocent, without fault
\se kebersihan
\ps n
\ge cleanliness
\se membersihkan
\ps vt
\sn 1
\ge clean_up
\de clean s.t. up
\sn 2
\ge purify
\de purify, repent or renounce
    immoral actions
\se pembersih
\ps n
\sn 1
\ge cleanser
\sn 2
\ge janitor
\dt 17/Jun/92
```

bersih *adj.* 1) be clean, not dirty or messy. 2) be innocent, without fault.

kebersihan *n.* cleanliness.

membersihkan *vt.* 1) clean s.t. up. 2) purify, repent or renounce immoral actions.

pembersih *n.* 1) cleanser. 2) janitor.

```
\lx bren
\ps vi
\ge play
\ee Implies lack of focus or
    purpose.
\se brenak
\ps vt
\ge play_s.t.
\de play a game, or play with
    s.t
\se inabren
\ps n
\ge recreation ; entertainment
\se rabrenak
\ps n
\ge toy
\dt 17/Jun/92
```

bren *vi.* play. Implies lack of focus or purpose.

brenak *vt.* play a game, or play with s.t.

inabren *n.* recreation, entertainment.

rabrenak *n.* toy.

Summary: The **\se** and **\ps** fields begin the new subsection of an entry at a new line. The **\sn** field continues on the same line. The *relative hierarchy* is as follows:

\lx lexeme
\ps part of speech
\sn sense number, \sn sense number
\ps part of speech
\sn sense number, \sn sense number, \sn sense number
\ps part of speech
\se subentry
\ps part of speech
\sn sense number, \sn sense number, \sn sense number
\ps part of speech
\sn sense number, \sn sense number
\se subentry
\ps part of speech
\sn sense number, \sn sense number
\ps part of speech
\se subentry
\ps part of speech
\ps part of speech
\sn sense number, \sn sense number

The **\lx** and **\ps** fields are the only ones that are minimally required for structuring entries (along with **\ge**, etc. to give useful information within the structural hierarchy of an entry). **\se** and **\sn** should only be used as they are appropriate for substructuring an entry.

2.5 Direct character formatting within a field

All fields are given a basic character style when printed. For example the **\ge** field is marked as being "English", the **\gn** field is "national language" character styles. Fields marked as "vernacular" include all of the cross-reference type fields **\cf**, **\sy**, **\an**, etc., as well as the obvious ones: **\lx**, **\se**, **\xv**, etc. Because the data within each of these fields are in a single language there is little problem in assigning character styles to them automatically. The contents of the entire field is given the same typeface. But the world is not so easy for "free-form" discussion type fields, and so MDF provides for direct character formatting in any field.

Although free-form fields are also given a basic character style (e.g., the **\ue** field is marked as "English"), they often contain words or phrases in the vernacular because they are designed for discussion of the vernacular language. This vernacular text is set off from surrounding information in a discussion field by preceding the vernacular word with the code **fv:** (for *font-vernacular*). The print tables use this code to apply the vernacular character style to the word that follows it.

How it is entered in the lexical database:

```
\ue The kin term fv:wai is
    used for ...
```

How it prints:

Usage: The kin term **wai** is used for ...

TIP: For this type of coding to work, *there must not be any space between the colon* (:) *and the following text* (this distinguishes the language code from normal punctuation), and the code must be in *lower case* (i.e. **fv:**, not FV: or Fv:).

Be sure to place the code *with* the word *inside* punctuation (parentheses, quotes, etc.). Otherwise the punctuation will receive the character style along with the word. For example, if you want to print: "*...during a hunt, the dogs (***asure***) go out ahead...*", the vernacular occurs in parentheses; encode this as "...dogs (fv:asure) go..." and *not* "...dogs fv:(asure) go..."

If the vernacular text is a *phrase*, the phrase should be linked together with an underline character: "using fv:mbwai_ka in most cases ..." The print tables would then apply the character style to the whole phrase, changing the underline character to a space in the process.[14]

The character styles do not flow across punctuation. Thus, character formatting codes must be placed on both sides of the punctuation. For example, **fv:peni/fv:beka** prints as "**peni/beka**", whereas **fv:peni/beka** prints as "**peni**/beka".

Character styles for other languages are set off as follows:

fn: for the national language (i.e. *font-national*)
fe: for English (i.e. *font-English*, if ever needed)
fr: for the local regional language (i.e. *font-regional*)

Other useful character styles are:

uc: (underline characters—see discussion below)
ui: (underline italic characters)
ub: (underline bold characters)
sc: (scientific name—set as underline italic, not required in **\sc** field)

The **uc:** code is able to detect which type of field it is used in. If the field is a vernacular field, **uc:** will underline with bold characters (following the vernacular character style); if the field is for the national language, **uc:** will underline italic characters; and if the field is for English, **uc:** will underline normal characters. If specific control is required, use **ui:** and **ub:**.

[14]Alternatively one could add an **fv:** before each word in the phrase, but this increases the typing load. Either way will work.

All of these codes are to be used in the same way as described for the **fv:** code.

To reiterate what was said above, character style codes are *unnecessary* in most fields because the field contains only one type of data (e.g. the national language gloss in the **\gn** field does not need to be marked as national language). Such fields are converted to the appropriate character style automatically. Direct character formatting codes are used only in general information fields or discussion free-form fields where language data and discussion are mixed.

> **TIP**: Use these codes to keep language styles consistent throughout your dictionary. Where possible, using the codes based on function (e.g. fv: fn: sc:) is preferable in the long-term over using the codes based on form (e.g. ub: ui: uc:). This function-based strategy facilitates uniform editorial changes and systematic upgrades to future generation computer software.

The use of the **uc:** underline code is very helpful in example sentences that focus on particles, functors, affixes, etc. In an Indonesian dictionary the entry **\lx di** might contain the example sentence **Bukunya tidak ditaruh di atas meja ini**. This is encoded:

```
\xv Bukunya tidak ditaruh uc:di atas meja ini.
```

So, even though the sentence has two "**di**" morphemes in it (**di-taruh** [verbal prefix] and **di** [preposition]), the underlining is used to mark the lexeme in question.

Underlining affixes often poses a problem. For example, if the third person singular pronoun possessive suffix is **-a**, it needs to be underlined in a sentence such as **Aulopoa aua lae weidu**, because there is another word that ends in '**a**'. But, because **–a** is only *part of a word*, underlining it with **uc:** will not work. To underline the '**a**' we must resort to the rather inelegant bar code and curly braces:

```
\xv Aulopo|u{a} aua lae weidu.
```

The **|u** marks the bracketed character as underlined and bold (a type of vernacular style). Note that these braces can be used to enclose any number of letters; this code is not restricted to use with just single letters.[15] When using this code *be sure to include the closing brace!!* If you forget it, the rest of your dictionary will be underlined! For this very reason the colon type of character style codes were developed. The bar code **|u{}** like **uc:** can determine what type of field it is in and adjust the underlining to match the surrounding character style.

[15]In fact, this is the general underlying form the code **un:** produces on the word and phrase level when the lexical file is being formatted for conversion over to a WORD document.

2.6 Punctuation

Leave off all punctuation *at the end* of straight data fields (**\ps**, **\ge**, **\cf**, etc.). The only places where punctuation should be included is in and at the end of free-form (discussion type) fields (**\ue**, **\ee**, **\nt**, etc.). All other field-final punctuation is added by the conversion process automatically.

For some national languages, such as French, there are orthographic conventions that encourage the use of special characters for punctuation. Some compilers use the chevrons « » in their SHOEBOX database to indicate double quotes for French and for the vernacular in French-speaking countries. However, MS-WORD reserves these characters for the macro language and the computer reacts to them differently than to other characters, giving messages and inserting asterisks in the text when importing the formatted file into WORD from MDF. We recommend using the Anglo-centric option of double-quote marks “ ”, which is an alternative punctuation convention for French. Once the file is imported into WORD, then the double-quotes can be replaced by chevrons if desired.

3. Introduction to the Multi-Dictionary Formatter program

This chapter documents the Multi-Dictionary Formatter, v1.0, December 1994. For changes from versions 0.9x see Appendix F.

The purpose of the MDF program is to assist you in structuring and formatting your vernacular dictionary and creating and formatting your English and national language "finderlists" (i.e. reversed listings of your vernacular dictionary).

NOTE: *The MDF program does not modify or in any way change your original lexical database*. Your database is simply read and the needed information extracted to another file where further processing is done.

CAUTION: If your lexical database does not use the standard field codes recognized by MDF, do not use this program yet. First convert your lexical field codes to this standard (as explained in chapter 2). This conversion only has to be done once and enables the user to tie into all of the formatting power and flexibility that MDF provides. Converting your codes can be done with a CC table or by using the EDIT REPLACE feature of WORD.

3.1 Familiarizing yourself with the program

First, test the way MDF is set up on your computer and how it interacts with your particular word processor by using MDF with the sample file provided on the release disk, called MDFSAMPL.DB. You can look at this file in SHOEBOX (or in a word processor if you do not make any changes and save it again as 'text only'), and then process it in MDF by using the following command:

```
C:\MDF>mdf mdfsampl.db<ENTER>
```

Try the '**F**ormat dictionary' and then '**E**nglish finderlist' options to become familiar with the various menu options MDF provides. Answer the questions prompted by MDF on the screen. The vernacular language in MDFSAMPL.DB is *Selaru* and the national language is *Indonesian*, but for becoming familiar with the program you can fill in whatever you like, including the vernacular language and national language appropriate to your situation. This database has also been formatted through MDF into a triglot dictionary with examples and notes (file MDFSAMPL.DOC on disk) for you to view directly through WORD. A formatted English reversed listing is also included (file MDFSAMPL.ENG). Together these will give you some idea of how MDF interacts with the database file to produce the formatted document.

Before you try out MDF on your full-sized lexical database, we recommend you make a sample database of about 40–50 records copied from your main database. (If you use WORD to do this, save the sample database as 'text only').[1] Run this sample database through MDF, selecting the different configurations available and saving the results to different filenames; and then print the different output files to see which format you like best. This suggestion applies to the formatted dictionary as well as to the national language and English finderlists.

3.2 Requirements and limitations

The current version (1.0) of MDF is set up for WORD-for-DOS v5.0, v5.5, or v6.0 and WORD-for-WINDOWS (WINWORD v2.0 and v6.0).[2] You will be asked to specify your word processor. In order to run, MDF needs to know the *full filename* of your lexical database. If the database is not in the MDF directory, include the path. For example, if LEXICON.DB is in the C:\SAWAI subdirectory, type:[3]

```
C:\MDF>mdf \sawai\lexicon.db
```

When MDF starts, it will ask you to specify the version of WORD you are using. (Use the arrow keys and <ENTER> to select it.) If you prefer to specify this from the command line, the following exemplifies how to do it:

```
C:\MDF>mdf lexicon.db v5        (for WORD v5.0)
C:\MDF>mdf lexicon.db v55       (for WORD v5.5)
C:\MDF>mdf lexicon.db v6        (for WORD v6.0)
C:\MDF>mdf lexicon.db win2      (for WINWORD v2.0)
C:\MDF>mdf lexicon.db win6      (for WINWORD v6.0)
```

The MDF program can have trouble merging documents in WORD v5.5 and WORD v6.0 simply because the glossary files used by those programs assume a default keyboard setup for each version of WORD. If the user has configured the keyboard in WORD to be different from the default configuration, MDF may malfunction at the point where WORD is called. So this is one reason we recommend testing MDF on a small section of

---

[1]Be sure to turn off automatic pagination and autosave *before* you load your lexicon. If you happen to alter the lexical file in any way, autosave will save a temporary copy of the file in WORD format (even though the file is text only) and this takes 'years' for large lexicon files! Auto-pagination inevitably slows the program down.

[2]If the user specifies WINWORD as the word processor, MDF will format, split, and convert the database files to WORD documents, but makes no attempt to merge them (because MDF cannot access WINWORD). The user will need to exit MDF and load each document file into WINWORD manually for merging and printing. For WINWORD, formatted dictionaries are named DICTN*.DOC, English reversed lists are ENGLS*.DOC, and national reversed lists are NATNL*.DOC.

[3]We are aware that there is some overlap between the material in this section and that in chapter 1. The overlap is intentional.

your lexicon to see that all is working well before trying to process your whole lexicon. If MDF does not work properly, exit MDF, reconfigure WORD to its default settings, and try MDF again.

Although most users will be quite pleased with the results, MDF is *not* a sophisticated program (from a computing point of view). It requires some user care. Be sure there is enough *free space* on the default drive to process your dictionary and finderlists. A safe size is at least four times the size of the original lexical database. This should give enough space for the working files as well as the final document files for the formatted dictionary and finderlists. Using MDF on a floppy drive would be unwise—it will probably not know when it has run out of room.

The MDF program reserves the filenames DICT*.*, ENGL*.*, and NATN*.* for its own use (to create the formatted dictionary, the English reversed list, and the national language reversed list, respectively) as well as SPLIT*.* for some working files. Please do *not* use these filenames for your own work (especially within the default directory where MDF resides). Files with these names will be deleted by MDF!

MDF must be able to find the MS-DOS program SORT.EXE. If it is unable to find SORT, it will not be able to run properly. To test if MDF will be able to find SORT, type DIR | SORT at the DOS prompt:

```
C:\DICT>dir | sort
```
[| = vertical bar, not colon]

If this gives an *alphabetized* listing of the files on the default directory (the "bytes free" line is also sorted to the top), then all is okay, but if the files are not sorted alphabetically, then the SORT program is not available. You will need to either specify a path that makes SORT accessible, or you will need to copy SORT to a place where it can be found (such as the directory where MDF and its associated files are).

MDF must also be able to find your word processor. MDF assumes that your word processor subdirectory is specified in the PATH command of your AUTOEXEC.BAT file and that your word processor is named WORD.EXE. If you have more than one version of WORD installed and have renamed the files (e.g. WORD5.EXE and WORD6.EXE), make sure the version you want to use with MDF is named (or renamed) to WORD.EXE. Make sure that particular subdirectory is added to the PATH command in AUTOEXEC.BAT. To check this, from the MDF subdirectory type:

```
C:\MDF>word<ENTER>
```
[check for WORD-for-DOS]

```
C:\MDF>win winword<ENTER>
```
[check for WORD-for-WINDOWS]

If your word processor comes up, then the setup is as it should be.

3.3 Overview of menu options

After specifying your word processor, MDF opens with the following menu:

Multi-Dictionary Formatter
Overview (shows you this chapter) **F**ormat Dictionary **E**nglish Finderlist **N**ational Finderlist **C**hange Settings **R**eset

Of the six choices here the first four are relatively transparent. The last two options require some exlanation and are addressed first.

3.3.1 Change Settings

The MDF program is set up "from the factory" to exclude certain lexical fields from the formatted vernacular dictionary. [**NOTE**: creating finderlists makes no use of these settings]. The excluded fields are:

\we	(word level gloss—English)	\is	(index of semantics)
\wn	(word level gloss—national)	\th	(thesaurus)
\wr	(word level gloss—regional)	\es	(etymology—source)
\re	(reverse—English)	\ec	(etymology—comment)
\rn	(reverse—national)	\so	(source)
\rr	(reverse—regional)	\st	(status)
\xg	(example glossing)	\dt	(datestamp)
\sd	(semantic domain)		

MDF also excludes all unknown fields (i.e. fields not found in the standard set given in the accompanying guidelines). These are coded in the settings file with '(huh)'. MDF by default also excludes all SHOEBOX created fields (_no, etc.). *All other fields are printed if present.*

The default settings can be modified either by excluding fields that would normally print or by including any of the above fields that normally would not print.

> **TIP**: Before using the 'Change Settings' option users should familiarize themselves with the built-in formatting options that MDF provides through answering a number of MDF-prompted options after selecting 'Format Dictionary' as explained below.

Selecting 'Change Settings' will call a simple text editor (TED.COM) and load a CC table file which you modify. How it is to be modified is explained in the file, but basically

you *add* a 'c' to the beginning of the line of any field you *don't* want to print, and *remove* the 'c' from the beginning of the line of any field you *do* want to print. (The 'c' means 'comment' or 'ignore'). Keeping things lined up is not important.

Save the file by exiting (F7–Exit) and <ENTER>. Your changes will be used to create a new settings file. Later when you want to format your vernacular dictionary, select 'Format Dictionary' from the menu. Your new settings will be used to create the formatted dictionary.

Before the dictionary formatting process begins, you have the following options:

1) Excluding example sentences (this would exclude the **\rf, \xv, \xe, \xn, \xr, \xg** fields)

2) Excluding your notes (this would exclude the **\nt, \np, \ng, \nd, \na, \ns,** and **\nq** fields).

These formatting choices supersede the settings file for discarding fields. But if the settings file is set to discard, say, the **\rf** field, choosing to *include* example sentence fields does *not* override the settings file and cause the **\rf** field to print. Only the **\xv, \xe, \xn,** and **\xr** fields would be output in this case. These options allow you to quickly alter an output format for a particular audience (e.g. the dictionary for a national audience would normally not contain your notes, whereas your own printed copy would), without having to go through the Change Settings menu option and mark each of the example sentence or note field codes to be ignored.

3.3.2 Reset

This menu choice simply restores the settings file back to its original 'from the factory' form. This resets which fields are excluded from the dictionary back to the ones listed above in §3.3.1.

3.3.3 Format Dictionary

> **NOTE**: For users of SHOEBOX v1.2x (and earlier), your database does *not* need to be "compacted" before using MDF. The file is resorted anyway to order homonym numbers correctly.

While formatting a dictionary in MDF is a fairly fast and automatic process, it is by no means simple. The following describes in more detail what actually goes on behind the scenes. Each of these steps is performed by MDF automatically and relatively quickly.

When MDF is processing your dictionary, it produces several intermediate files, but without altering your original lexical database. The first step is to throw out every field

that you have specified in the settings file that you do not want (see §3.3.1). MDF puts a dot on the screen for every record it processes.

The output file is then sorted, taking into consideration homonym numbers.[4] This second step is necessary because SHOEBOX sorts only on the KEY field contents. With homonyms, key fields are identical (see §6.3), and SHOEBOX assumes therefore that there is no particular order for such records. In fact, SHOEBOX reverses the order of homonyms each time it compacts the file. So there is no point in worrying about keeping the homonym records in numerical order—you just can't.

Now since homonyms are marked as 1, 2, 3, etc. and it would look rather odd to have sets of homonym entries printed in random orders, MDF sorts them on both the **\lx** field and the **\hm** field. (see also §5.4.1).

This sorting process uses the Text Analysis [TA] program SRT.EXE supplied with the MDF release. The default sort order is in the file MDFDICT.ANS. The sort order may be modified (outside of MDF) using the TA program ANSQ.EXE (this will be important for users with digraphs or other complex orthographic issues). Changing the sort order is explained in the documentation that comes with the ANSQ.EXE program. An alternative means of changing the sort order in MDFDICT.ANS is explained in §5.4.2. But, for MDF to function properly, the @ symbol *must be sorted first*. This symbol is used to sort a dummy record to the beginning of the sorted file. This first record contains setting information used by MDF later in the formatting process. This extra record also causes SRT to give a record total that is one greater than the actual number in your lexical database. For MDF to function properly, the MDFDICT.ANS file must contain the line that tells SRT to use both the **\lx** and the **\hm** fields when sorting:

```
\rkey lx hm
```

Once sorted, the database file is then processed by a large CC table to convert it to a file with all of the necessary paragraph and character style codes assigned to the appropriate bits of text, with new letter sections added, with odd-even running footers, and all of the other things necessary to get it ready for moving over to WORD.

The output of this CC table is then split into smaller, more manageable files, called SPLIT01.TMP, SPLIT02.TMP, etc. These are then input into the Convert-to-Word [CTW] program one by one.[5] The CTW program then does some serious crunching on the files to produce a series of printer-ready WORD documents (still in pieces). These

---

[4]The homonym number applies to the entry citation form if there is no **\lc** field. Then the **\hm** number applies to the **\lx** form if there is a **\lc** field present. Then the **\hm** number references the **\lc** field, *not* the **\lx** field. The user must keep this distinction in mind.

[5]CTW is a good program, but because it is limited in the size of the input and output file, the database file must first be split into smaller files.

document files are called SPLIT01.DOC, SPLIT02.DOC, etc. and they must then be merged back together in WORD.

The final step loads a sorted list of the split document files into WORD. This list is used to remerge the files. The merged document is then loaded into WORD for your perusal. This file is given the temporary filename MDFXXX.TMP. After the file has been viewed, simply quit WORD, and MDF will change the temporary name to the name DICT.DOC. (You will be notified of the new name by MDF). If you wish, you can rename the MDFXXX.TMP file to something else from within WORD (v5.0 use TRANSFER–RENAME; v5.5 or v6.0 use SAVE AS). Renaming the file will not affect the MDF program. It assumes that if MDFXXX.TMP no longer exists, you must have already given it another name.

Once merged, the dictionary is basically ready for printing (though you may desire to make cosmetic changes). This process from a standard format lexical database to a printer-ready document is relatively automatic. It takes MDF about 13 minutes to format a vernacular-English diglot dictionary from a 791K lexical database with 2,044 records (many of them complex) on a Toshiba T1900 laptop (a 486SX–20MHz machine). It takes over 45 minutes on a PC–XT. The following example illustrates a triglot printout.

Sample SHOEBOX Records

```
\lx abat
\ps n
\ge grove
\gn dusun
\rf d2.077.03
\xv Kbwai abatke ti ksweruk
    nurare.
\xe I went to the coconut
    groves to clear the grass.
\xn Saya pergi menyiangi dusun
    kelapa.
\rf d4.079.16
\xv Kbwa ti ktwan nurke o
    abatke.
\xe I'm going to plant coconut
    trees in the grove.
\xn Saya pergi tanam kelapa
    di dusun.
\ee This is uc:not limited to
    coconut groves but is used
    for mangoes, etc.
\sg abatke
\pl
\nt
\dt 26/Feb/90
```

MDF Triglot Output

abat *n.* grove; *dusun*. *Ref:* d2.077.03 **Kbwai abatke ti ksweruk nurare.** I went to the coconut groves to clear the grass. *Saya pergi menyiangi dusun kelapa. Ref:* d4.079.16 **Kbwa ti ktwan nurke o abatke.** I'm going to plant coconut trees in the grove. *Saya pergi tanam kelapa di dusun.* This is not limited to coconut groves but is used for mangoes, etc. *Sg:* **abatke**.

```
\lx -abili
\ps v
\ge wail
\gn meratap
\rn ratap, me-*
\rf n2.113.30
\xv Kswer ma kabili yaw ti
    lasmyerke.
\xe I wailed prostrate on the
    ground.
\xn Saya meratap di tanah.
\cf -ser
\ce cry
\cn menangis
\pd 1
\1s kabili
\nt
\dt 1/Feb/90
```

-abili *v.* wail; *meratap*. *Ref:* n2.113.30 **Kswer ma kabili yaw ti lasmyerke.** I wailed prostrate on the ground. *Saya meratap di tanah. See:* **-ser** 'cry' '*menangis*'. *Prdm:* 1. *1s:* **kabili**.

3.3.4 English and national language finderlists

Some commercial bilingual and trilingual dictionaries are quite detailed in their description of each language. The good ones are really two separate dictionaries from different perspectives (language 1 as expressed in language 2, and language 2 as expressed in language 1—which are rarely reciprocal). Such complementary dictionaries can be produced in SHOEBOX and MDF through two separate databases. But most field researchers can invest heavily in only one point of reference (vernacular to English and/or the national language). Dictionaries based on field research are not normally intended to explain English or the national language to the local language group, but to provide a detailed inventory of the local language and make this accessible to outsiders. (See §4.1, §4.2, and §4.3 for related issues.)

For most field researchers, a reversed index (or finderlist) will be sufficient. These finderlists provide the needed links from English or the national language to the local language. A term referenced in a finderlist can be found in the main dictionary should the user need a more detailed explanation of the term.

MDF produces formatted national language and English finderlists by making two separate passes through the lexical database (one pass for each list). A finderlist is produced as follows:

First, the lexical database is processed with a CC table to extract and reverse the glosses. This produces an unsorted file. The unsorted file is then sorted using the SRT program and the sort specifications found in the file MDFENGL.ANS or MDFNATN.ANS (for the English or the national language lists, respectively). The sorted output is then processed by another CC table to collapse (merge) identical English or national language

entries into single entries. This collapsed database file is now ready for processing through another CC table to become a formatted file ready for conversion to a WORD document. The program CTW (which does the converting) is unable to handle large files. So the formatted file is split into smaller files, as is also the case when formatting the dictionary. These are then run through CTW one at a time. Finally a list of these split files is loaded into WORD, and WORD uses the list to merge the split document files back into a single document. This produces a printer-ready document in WORD.

The document files are merged into a temporary file called MDFXXX.TMP. The user is given a chance to look at the finderlist while it is still called this. It may be renamed if needed (in WORD v5.0 using TRANSFER–RENAME; in WORD v5.5 or v6.0 use SAVE AS). If you choose not to give it a new name, exit WORD, and the new finderlist is automatically given the name ENGL.DOC or NATN.DOC depending on which language it is for.

This whole process must then be repeated to produce a finderlist for the other language. On a 486SX 20Mhz laptop, MDF takes just over five minutes to produce an English finderlist from a 791K lexical database, with 2,044 records.

The essence of making a reversed finderlist involves storing the lexical entry form, the lexical citation form (if present), and the subentry form (if there is one), as well as the homonym number and the current sense number (if relevant), and then outputting a reversed record for each gloss occurring for the language being extracted.

The finderlists produced can be in either single or double column format and can either include or exclude the part of speech of the vernacular term being referenced. The following examples are single column:

With the part of speech:

enrage	*adj*. **masbu**.
enter	*vi*. **-sukar**.
entertain	*vt*. **-aluka**.
entertainment	*n*. **inabrenke**, see: **-bren**;
entire	*vi*. **ktem**$_1$.
envious	*ph*. **lema kdwakin irire wait eraske**, see: **-dakin**.
erase	*vt*. **–sos**.

Without the part of speech:

enrage	**masbu**.
enter	**-sukar**.
entertain	**-aluka**.
entertainment	**inabrenke**, see: **-bren**;

entire **ktem**$_1$.
envious **lema kdwakin irire wait eraske**, see: **-dakin**.
erase **–sos**.

The MDF program combines the vernacular glosses in identical reversed glosses (shown below with the part of speech). (Note with long headwords MDF pushes the part of speech and gloss further to the right on reversal so that only the shorter units are fully aligned.)

face *n:bp.* **mata**;
n:bp. **welnohaha**.
face, to wash one's *vi.* **-larif**.
faded *adj.* **mamwaw**.
faithful *vi.* **-tohtohaktel**.
fake *adv.* **koikay**.
falcon *n:an.* **lak**.
fall *v.* **kibrok**;
v. **-tunik**;
vi. **-di**;
vi. **kdi**;
vi. **kyoras**;
— kdian.
fall forward *v.* **-surak**.

The same list is shown below without the part of speech (Note that multiple references, such as 'fall', are concatenated sequentially rather than displayed on separate lines as above):

face **mata**; **welnohaha**.
face, to wash one's **-larif**.
faded **mamwaw**.
faithful **-tohtohaktel**.
fake **koikay**.
falcon **lak**.
fall **kibrok**; **-tunik**; **-di**; **kdi**; **kyoras**; **kdian**.
fall forward **-surak**.

The total number of entries in each finderlist is given as a statement at the end of the document.

3.3.5 Quit

To leave MDF hit the <ESC> key at the main menu. A message giving the version and date of the MDF program will be displayed as it returns you to DOS.

You are now free to reload each of your document files (DICT.DOC, ENGL.DOC, NATN.DOC) into WORD to tweak as needed (margins, headers, footers, etc.). If you find errors in the actual text due to MDF please report them using Appendix I. If you find errors due to your own mistakes in the lexical database, you can go ahead and correct them in the printer-ready dictionary, *just be sure to also correct the errors in the original lexical database*; otherwise you will have to correct those errors every time you format your dictionary.

3.4 Printing

The MDF program was designed to get everything ready for printing, but not to actually handle the printing.

Once your dictionary and finderlists have been formatted, exit MDF and then use WORD directly to load and print them. Or you could print them from within MDF right after each document is merged into WORD.

Before printing a large print job, first print a couple of pages to check that the interaction of the stylesheet with your printer is satisfactory. Several stylesheets are provided on the release disk as explained below. Select (and if necessary adapt) the stylesheet that is most appropriate for your printer.

If you are printing your dictionary on a dot-matrix printer (or perhaps on a light duty inkjet printer), have WORD print only 20 or so pages at a time. Let the printer rest a bit and then continue. This helps keep the print head from overheating. Another solution is to open the lid and direct a fan at the print head. This may allow you to print the whole file at one pass.

The stylesheet MDFDICT.STY is automatically attached to each of the final documents by MDF. It is set up for the HP Laserjet series printers (III and above; the file MDF-HP4L.STY is identical to MDFDICT.STY). It also does a fairly nice job for the Epson LQ series printers (though the MDF-EPLQ.STY stylesheet is designed for these printers). MDFDICT.STY bombs on the Toshiba 321SL, so if this is your printer, you will need to copy the stylesheet MDF-T321.STY over to MDFDICT.STY so that MDF will attach a "Toshiba 321SL" version of MDFDICT.STY to each document.

If you want to modify the look of your dictionary and finderlists, modify MDFDICT.STY, but be sure to also save the modified version to another filename, such as MY.STY. This allows you to switch to other printers (by copying another printer style over MDFDICT.STY) and not lose all the modifications you made for your own printer (just copy your stylesheet back to MDFDICT.STY when you want to use it again).

There is also a stylesheet called MDF-FLIP.STY which "flips" your document from a single-column format to a double-column one, or vice versa. So even if you choose

"double-column" format when MDF asks you, you are not stuck with the decision, just attach MDF-FLIP.STY and the document is automatically changed. Reattaching MDFDICT.STY returns the document to the original format.

MDF-FLIP.STY is a modification of MDFDICT.STY (it is identical to MDF-HP4F.STY), so if you are using a modified stylesheet like MDF-T321.STY or one you've made yourself, you will need modify MDF-FLIP.STY too, if you want to use it. Again be sure to also give the modified stylesheet a new name, such as MYFLIP.STY.

3.5 Modifying the printout

3.5.1 WORD Stylesheets

The easiest way to modify the look of your formatted dictionary and finderlists is to modify the WORD stylesheet MDFDICT.STY, giving it a new name after you've modified it. This stylesheet is used by both the dictionary *and* the finderlists, so beware: what you do for the dictionary may affect the finderlists as well. If it does and you don't like it, then make two stylesheets, one for your dictionary and one for the finderlists. You will need to attach your modified stylesheets *each time you want to print*. MDF does not know about them and stubbornly attaches MDFDICT.STY to the documents.

Most of the styles in the stylesheet are *character styles*. It is pretty clear for most styles what they affect (e.g. SN style formats the sense number, etc.). But the FV, FE, FN, and FR styles affect more than just one lexical field. These codes (for vernacular, English, national, and regional fonts respectively) determine the look of most of the fields. These styles are used for all language specific text (**\dv, \de, \dn**, etc.). So, for example, if you print out a diglot dictionary for a national language audience, you will probably want to tweak the FN style, because this style is set to italic (to differentiate the national language from English in a triglot dictionary). Simply edit the stylesheet and change FN back to normal text for your national diglot dictionary.

The standard font style [FS] is used for formatting most information fields (**\rf, \lt, \pd, \lf, \is, \th, \sd, \bw, \et,** and **\cf**), as well as for punctuation.

The *labels* that mark different fields (e.g. *See:* for the cross-reference field) are all encoded with the FL style (mnemonic for "font—label").

3.5.2 Character Style codes

MDF supports embedded coding in your discussion fields so that you can apply or specify a character style to any bit of text in your dictionary. These embedded codes are to be used in your lexical database before the dictionary is formatted, not afterwards. The following are the character style codes supported by MDF (see also §2.5):

fv: (font—vernacular)
fe: (font—English)
fn: (font—national language)
fr: (font—regional language)

fl: (font—labels)
fs: (font—standard)
fb: (font—bold)
fi: (font—italic)

uc: (underline character)
ub: (underline—bold)
ui: (underline—italic)
sc: (underline a scientific name—not required in the \sc field)

These codes can be specified within any field (but generally are used only in free-form or discussion fields). When specified, they apply to the following word (a space or punctuation terminates the style). The style codes must be in lower case, and must not have any space between the colon and the following word:

```
\ee They make fv:sabun using pulverized coral...
```

This would print as:

They make **sabun** using pulverized coral...

Use the underline (**_**) character to link words in a phrase: e.g. **fv:bikin_apa_di_sini?** To mark the character style of a word inside parentheses, quotes, brackets, etc., the character codes must be placed with the word *inside* the enclosing punctuation.

The **uc:** code is able to detect which type of field it is used in. If the field is a vernacular field, **uc:** will underline with bold characters (following the vernacular character style); if the field is for the national language, **uc:** will underline italic characters; and if the field is for English, **uc:** will underline normal characters. In order to specifically control the underlining character style, use **ui:** and **ub:**.

After the dictionary is formatted, if you find you missed some piece of text that needs a character style code, simply use the same letter codes as if applying a style in WORD (e.g. ALT+F,V for vernacular font in WORD v5.0, or CTRL+SHIFT+F,V for vernacular font in WORD v6.0, etc.). But remember to go back to your lexicon and also put in the vernacular font code (**fv:**) where needed (so you won't have to add it every time you format your dictionary).

3.6 Summary

We hope the MDF program makes the whole process of printing out your dictionary and finderlists easy enough so that it can be done as frequently as needed. In printed form, a dictionary can be a valuable language learning tool for you, helpful to others in related languages, and also a good demonstration of progress to the language community and the government authorities. A dictionary only on the computer is of little use to anybody but yourself (and then only when you are sitting at the computer).

4. Basic strategies and perspectives

Several preliminary issues discussed here will help with understanding the bigger picture in dictionary-making and in choosing between different strategies.

4.1 Terminology

Lexicon$_1$: We use the term *lexicon* in two different ways in this *Guide*. In the linguistic sense it is the *vocabulary* of a language, including compounds, idioms, and other phrasal units.

Lexicon$_2$: In the data management sense *lexicon* is used to refer to the lexical *database*; the physical inventory of the lexicon$_1$. It includes additional information and coding related to cross-referencing, reversal, formatting, and housekeeping.

Dictionary: A restricted portion of the lexical database (lexicon$_2$) that is published for a primary purpose and a primary audience. A dictionary provides a systematic exploration of the vocabulary of a language, including, among other things, meaning, range and usage. A dictionary normally uses some convention of alphabetizing to organize the material. Dictionaries normally do not include housekeeping information, but extract information from the lexical database for formatting. The broadest kind of dictionary is a comprehensive general purpose monolingual or bilingual dictionary. More specialized dictionaries might focus on kin terms, body parts, plants, fish, or animals. A medium-sized dictionary for publication has around 5,000 entries. A significant dictionary has over 10,000 entries (counting headwords as an entry).[1]

Glossary: A glossary is usually no more than a listing of the headword (lexeme) and a simple gloss or two. Sometimes it also includes part of speech. It does not include example sentences, synonyms, multiple senses, etc. A glossary is sometimes a necessary minimum for archiving dying languages and cultures, but should not be the goal or final result of any significant fieldwork. Minimal entries in a dictionary, and typical entries in a glossary are about the same.

Finderlist: A finderlist is similar to a glossary, but functions more like an *index* or a list to find vernacular forms that may sometimes be translation equivalents to the English or some other language. It could be seen simply as a list to find a form, without additional information. These are most often found as simple reversals of bilingual dictionaries, or

[1]Many commercial dictionaries count each separate part of speech, subentry, inflected forms, run-on derivatives and other classes of subsidiary information as separate entries for the purposes of inflating the total entry count. Thus, a single headword can be counted as five or more 'entries', because for commercial purposes the more entries one can claim, the more impressive (see Landau 1989:84-87). For the discussion in this *Guide* 'entries' are counted by headwords.

as simple dictionaries for comparative purposes within a family of related languages. MDF uses the term 'finderlist' for the various reversal options.

Thesaurus: A thesaurus is organized along different principles than a dictionary, generally around semantic domains. Very few general thesauruses for minority languages have been usable by the local communities. This is for a variety of reasons which are not yet well understood, but they include: the organizing categories chosen by the compiler do not fully match the categories recognized by the community themselves; and how to use a thesaurus is not immediately transparent,[2] etc. An attempt at a published thesaurus for a language is not recommended until a full dictionary has been published first.

However, a subset of the lexicon selected along semantic lines can be published prior to the publication of a major dictionary. This is best done as a selection of entries representing a generic term in the vernacular. The generic term *bird* in English covers a different range than the generic term *manut* in Buru. *Manut* (from Proto Austronesian ***manuk** 'bird') encompasses flying creatures whose wings are easily distinguished, including birds, bats, and butterflies, but not other flying creatures normally covered by the English generic term *insect*. Thus, a separate volume about *manut* could be published prior to the publication of the Buru dictionary, providing reading material and stimulating community interest. Similar volumes could focus on fish, animals, insects, reptiles, edible plants, jungle plants, kin terms, body parts, disease and medicines, etc. The **\th** field in MDF provides a place to record the vernacular generic term for later extraction or analysis.

4.2 Identifying the primary audience and purpose

A major issue which influences how other decisions are made is to have a clear idea about for whom the information is being packaged. The audience for a dictionary is usually one of the following:

The scholar/compiler: This is the default audience in which information is packaged for one's own convenience, reflecting the lowest level of thinking and organizing. It also is the audience that makes the information most difficult for anyone else to use. The compiler will generally know more than is put in the lexical database, simply using the database as a receptacle of cursory tags to jog the memory and organize information.

Academic audience: For an academic audience the compiler tends to use abstract terms, technical jargon, and occasionally even algebraic-like formulae (e.g. x **DO**–*cut* y *with* z, **CAUSE** y **BECOME** y'). A dictionary geared primarily to a linguistic academic audience

[2]Many educated westerners also have difficulty using thesauruses in major languages. How to find the information one is after often takes a larger investment of understanding than does using a dictionary.

tends to be fairly useless to other audiences, and is often used with great difficulty by other academics, if at all.

National government: A dictionary geared to please the national government often appears incomplete and full of shortcuts. It is often produced to justify a visa, on-going presence in the area, or show that contractual obligations are being met. It is rarely a service to anybody. A better option is to produce a serious volume for an academic audience that would contribute to both the local and scientific communities, and would deal with the visa or contract problems as well.

Local government: Local government officials with a variety of motives are frequently interested in a 'dictionary' to help them grapple with the local vernacular. What they usually mean is a simple glossary. However, the local community may not want the transitory civil service, police, or military to know certain areas of vocabulary, such as female body part terms and sexual terminology, and may request that certain areas of vocabulary be left out of something made for local officials. The information that will satisfy the needs of local officials is less than that required for a serious dictionary, and so is not recommended as a primary audience.

Local audience: The local audience often has a variety of purposes or desires in having a dictionary of their language. Prestige and ethnolinguistic pride may enter in—upon getting a dictionary it is not uncommon to hear, "Now we have a real language!" Community leaders may feel the younger generations are rapidly shifting to a regional or national language and want a reliable inventory of their language and culture in the form of a dictionary. Or they may feel that knowledge of certain parts of their language and culture (such as ritual language or traditional medicine) are not being transmitted to a new generation of specialists and need to be archived while the knowledge is still available. *The information catalogued in a serious attempt to make a dictionary that will serve the broad needs of the local community will normally serve other audiences and purposes as well.*

General audience: This is commonly cited as the primary audience of compilers of dictionaries. However, a 'general' audience is simply not specific enough to assist in decision-making about how information should be packaged or what information should or should not be included. A product aimed for a 'general' audience is often amorphous and unprincipled.

Mixed audience: This may be either something to be studiously avoided, or a viable solution to several problems. Trying to serve mixed audiences with mixed purposes can make a dictionary very unsatisfying or very unwieldy. For example, a dictionary geared primarily for an academic audience will probably not be usable by the local community. One solution is to make separate dictionaries for separate audiences. However, few

scholars have the time or the financial resources to make more than one serious dictionary.

OUR RECOMMENDATION: Given the reality that a compiler will probably be limited to producing one or at most two dictionaries, we recommend that the major dictionary be aimed for the local audience and supplement it with information that is of use to secondary audiences, such as a scholarly audience. For example, the addition of scientific names, etymological information, and morphological parsing (e.g. **memukuli** → meN-pukul-i) can nicely broaden a dictionary otherwise geared for a local audience. A viable solution is thus to aim the primary organization of the lexical information for a local audience, but to also embellish the entries with information that is useful and interesting to an academic audience. A well-organized computerized lexical database can accommodate information packaged for different audiences. The following example is from Buru:

`\lx sira`	[lexeme / headword]
`\ps PRO`	[part of speech]
`\ge 3p`	[gloss for interlinearizing texts]
`\gn mereka`	[gloss for national language]
`\gr dorang ; dong`	[glosses for regional language]
`\re they ; them`	[glosses for reversed English finderlist]
`\dv gebaro dikat fi di kita`	[definition—vernacular]
`\de they; third person plural`	[definition/description -English]
`\dn orang ketiga jamak`	[national language definition]
`\et *siDa`	[historical etymology]
`\eg they`	[gloss of etymology]

4.3 Monolingual, bilingual, and trilingual dictionaries

The purposes and organization of monolingual, bilingual and trilingual dictionaries vary greatly. A *monolingual* dictionary attempts to use the language to capture the essence and range of meaning and usage in such a way that the foreign, young, uneducated, or semi-proficient can understand and use a term. Definitions are of utmost importance, and must comply with rigorous technical and theoretical principles (see Wierzbicka 1992). They are very difficult to get right. Well-chosen examples can help reduce the complexity of technical definitions—carrying some of the weight, so to speak. Monolingual dictionaries are not the focus here, although the fields needed are supported by MDF and are discussed in this *Guide*.

A *bilingual* dictionary, focuses on providing translation equivalents (here called ‘glosses’) with reference to another language. The trick is to provide enough information so the user knows which glosses are appropriate (and inappropriate) in particular

contexts. Judicious use of examples assists with both justifying and exemplifying usage.[3] MDF provides for both vernacular-English and vernacular-national language diglot options. Pawley (1993:18/3/93 lecture notes) explains his view:

> In a bilingual dictionary, the situation is different [from a monolingual dictionary]. The bilingual dictionary, going from L1 to L2, is chiefly a translation aid and ideally it should be backed by monolingual dictionaries of the two languages. The user is looking for equivalents rather than analysis. Start with the ideal simplest case, where the two languages, L1 and L2, always have fully intertranslatable terms. By this I mean that for every term in L1 there is at least one term of equivalent meaning in L2. In such circumstances, the counterpart of the definition is the translation equivalent. And the lexicographer's job would be to specify the proper translation equivalent(s). There would be no need to define the meanings of terms in L2 analytically in the bilingual dictionary because the speaker of L1 would either know the equivalent term in his own language, or having been told it, would be able to look it up in a monolingual dictionary.
>
> However, bilingual dictionaries do not always work this way. The main reason is that the lexicons of different languages are never completely isomorphic—their semantic categories do not match one-to-one. Languages stemming from a common ancestor and spoken by communities with very similar cultures may show a fairly close match. So, sometimes, do unrelated languages whose speakers have been bilingual and in close contact for many centuries. But languages associated with radically different cultures may not be readily intertranslatable. Far from it. *In such cases, the lexicographer is obliged to give analytic definitions, in other words, to do much the same thing as the compiler of a monolingual dictionary.* Those of us who work on 'exotic' languages (from the European standpoint), such as Australian, Austronesian or Papuan languages, constantly find ourselves in this last situation. [emphasis added].

A *trilingual* dictionary (e.g. vernacular-English-national language) is visually cluttered and a nuisance to some users, but appreciated by others. Such a dictionary is generally not recommended for publication, although some communities feel they gain prestige by having the English along with the national language. If done at all, the decision to print a dictionary in trilingual format at the insistence of the local community should occur only after other alternatives have been fully discussed. It is generally better for the various audiences if the lexical database is divided into separate sections or even separate publications (i.e. vernacular-English; vernacular-national language). A triglot format is useful during the drafting and pre-publication stages to check for consistency and

[3]Most handheld electronic 'bilingual dictionaries' do not qualify as dictionaries in the sense used here. They are electronic *glossaries* (with varying degrees of sophistication). Multilingual dictionaries (e.g. eight European languages in a single volume) also tend to be glossaries without enough information to distinguish appropriate usage.

completeness. MDF provides for a vernacular-English-national language triglot option for this latter purpose.

4.4 Text-based lexicography and lexical sets of similar words

Pawley (1993:6/4/93 lecture notes) provides a preliminary context for compiling a dictionary:

> You can safely assume every language has at least 10,000 lexemes. If you are coming fresh to an exotic language how do you find the lexemes? There are several data-gathering methods.
>
> The most valuable thing you can do of course is to learn the language and culture. I doubt if a good dictionary can be compiled by anyone who does not have a reasonably good working knowledge of the target language and associated culture. But this takes time and you may want to start collecting immediately.

There are a variety of strategies for finding words to go into a lexical database.

1) What words can I think of?

2) What words do I know beginning with the letter *a*, for example?

3) Given the phonemes and phonotactic patterns of the language, what are the logically possible combinations of letters and morphemes, and which ones do the native speakers recognize as words?

4) Are there native speakers I can commission to fill in wordlists or think of terms for me? This approach is full of inherent pitfalls. These include: in many societies native speakers often have an inadequate mastery of the national language in which they try to describe or define the terms; there is mostly likely a mismatch of terms used in L1 and L2 even though the description is written by the native speaker on the assumption of a complete match; the compilers add further changes when they reinterpret into English what they are given, etc.

Strategies 1–4 are not recommended as primary (or serious) approaches. Some that have a little more merit include:

5) Are there good (tested) extended wordlists in the national language or a lingua franca I can use to get started? These are best if they are designed specifically for the language family. Because second (or third) languages tend to be used only in certain contexts or domains, be aware that there may be large areas of vocabulary that native speakers never use and may not know in the language of elicitation (such as the national language), but only in the vernacular (e.g. *centipede* or *leach*).

6) Is there a (good) dictionary of a related language that I can use to elicit forms and compare range of meaning? Here the compiler must take great cautions to avoid assumptions of isomorphism (one-to-one relationships of form and meaning across languages).

7) Are there good picture books (drawings or photographs) that can be used to elicit terms? They may be useful for flora, fauna, and material culture such as artifacts. However, there is the temptation to assume that the scientific name in the picture book is a perfect match for the native term, whereas the local varieties may, in fact, be different. Furthermore, scientific nomenclature often changes over time as botanists and zoologists refine the principles by which things are classified. Thus, the scientific name given by a qualified naturalist in 1850 or 1930 may not be what is used today, and what was covered by the term 100 years ago may be split into two or more terms now, or may have been merged with another term.

More sound approaches include:

8) What words relate to the semantic domain of *plants*, for example?

9) What words occur in a large corpus of natural texts?

10) How do sets of similar words compare and contrast in meaning?

The text-based strategy (approach (9), above) of looking for lexemes that occur in natural texts, forms a solid basis for building a good lexical database. While it should not be used as an exclusive strategy, it is highly productive and reliable as a primary source of words, and as a source for checking senses and investigating semantic and grammatical collocations. The computer program SHOEBOX is admirably set up for working through texts, automatically checking if words in the text are already in the lexical database, inserting them if they are not, and assisting the user to expand information in lexical entries.[4]

A caution for those building a lexicon primarily through interlinearizing texts: morpheme-level glossing for interlinearizing tends to encourage the compiler to ignore compounds and phrasal lexemes, and to overlook sense discrimination. After interlinearizing a text, parsing it by morphemes, it is wise to do a second pass through the text to identify polymorphemic words, compounds, and phrases that should be entered into the dictionary as separate headwords. Consider how English lexemes such as *book-keep-ing, short-stop,*

[4]A program called IT (Simons and Versaw 1987) is available to Apple MacIntosh users, but it does not have the extensive interactive capabilities available in SHOEBOX. IT works at the level of a glossary, rather than a full-blown lexical database. IT can be ordered from Academic Computing, 7500 W. Camp Wisdom Rd. Dallas, TX 75236, USA.

lawn-mow-er, touch-and-go, break-ing and enter-ing would be overlooked by morpheme-level interlinearizing.

To supplement text-based lexicography it is helpful to select headwords (lexemes) that are related to a single semantic domain (e.g. plants, houses, activities, emotions, etc.) and then compare and contrast a subset of similar terms within them (approaches (4) and (6), above). This is best done with one or more skilled native speakers.

For example, in Buru *kasa* might be glossed in isolation as 'roof rafter'. But in comparing *kasa* with terms for other kinds of roof rafters it becomes clear that *kasa* is limited to a certain function, a certain spatial orientation, and is normally only made from a couple of types of material, in contrast with other types of 'roof rafters'. Likewise, comparing and contrasting sets of similar lexemes like *trick, deceive, lie, tease, pull one's leg*, or sets of cutting verbs, carrying verbs, emotion words, or speech-act verbs enables the compiler to be precise and explicate the information salient to that particular term. *Thus, lexicography that explores lexical sets of related or similar words is more precise than that done in isolation.* Such information is often simply overlooked when the terms are considered in isolation. MDF provides the **\lf, \cf, \de, \ee, \ue,** and **\oe** field bundles for cataloging and linking the similarities and differences within lexical networks.

Pawley (1993:7/4/93 lecture notes) provides additional perspective and a caution:

> Learn the language. The best thing anyone wishing to make a first dictionary of a language is to become proficient in the language, gaining a good working command of the core vocabulary. Of course it takes years to learn a language well, to the point where you are familiar with the several thousand lexemes that are the stuff of everyday discourse. There are many quicker ways to gather data but without a first-hand knowledge it is difficult to evaluate data obtained by rapid methods. In my first spell of 11 weeks of fieldwork on Waya Island, Western Fiji, I used a rapid method that enabled me to record, after a fashion, perhaps 6000 word-forms and meanings that I believed to be Wayan. I started on the dictionary without yet having much knowledge of the Wayan language though I did have a working knowledge of Standard Fijian (the relationship is like English and Dutch or perhaps Dutch and German). It took me about 10 years to weed out all the mistakes in those 11 weeks. So much for fast track dictionaries.

4.5 Minimal entries vs. expanded entries

Another useful notion for those just beginning to compile a lexicon is the difference between minimal entries vs. expanded entries. In a computerized lexical database a minimal entry can always be expanded or changed when more information becomes available. For some purposes a minimal entry might simply include the word (lexeme) and a gloss:

\lx ama \ge father

ama father.

\lx ina \ge mother

ina mother.

Some compilers include housekeeping information in a minimal entry such as the date the entry was last worked on:[5]

\lx ama \ge father \dt 9/Sep/90

ama father.

\lx ina \ge mother \dt 8/Aug/89

ina mother.

Some who use the lexical database for linguistic analysis in interlinearizing texts want the part of speech included in a minimal entry:[6]

\lx ama \ps n \ge father \dt 9/Sep/90

ama *n.* father.

\lx ina \ps n \ge mother \dt 8/Aug/89

ina *n.* mother.

TIP: Fields you want to appear in every entry can be entered in the DATABASE TEMPLATE in SHOEBOX. (In version 2.0 this is found under FILE OPTIONS.) SHOEBOX will then insert these field markers automatically in each new entry. Remember to insert a space after each field marker in the DATABASE TEMPLATE.[7]

\ps· \ge· \dt·

Simple database template

[5]This is updated automatically in SHOEBOX if the DATESTAMP feature is enabled.

[6]However, there are good reasons to delay assigning part of speech, discussed in chapter 9. A common problem is that analysts tend to believe the labels that they assigned early in their exposure to a language before they understood how the language works as a *system*. We recommend that compilers flag tentative parts of speech assigned early in the language project in some way, perhaps with a preceding asterisk to indicate the tentative or hypothetical nature (\ps *vi). This will facilitate checks and later modifications once the system is better understood.

[7]To make sure there is a space at the end of each line, press <END> and check where the cursor sits.

Novice compilers will find it helpful to include many fields in their template—even more than they feel they need at the beginning. Empty fields are not a problem with MDF—if there is no content in a field, MDF ignores it when formatting the dictionary or reversed finderlist. By including many fields in the template, users will find the fields are there when they need them. Power users can add bundles of fields at any time using MACROS or direct keyboarding, but this is daunting to the beginner who is facing information overload. Fields will be consistent if entered by a template rather than by hand. We have a tendency to be lazy—if the field is *not* there we may fail to add the information even when we know it, but the *presence* of a field serves as a prompt. The following is suggested as a basic set of fields to include in every record in the lexicon. It is most easily entered in SHOEBOX as a DATABASE TEMPLATE.

\lx·	[lexeme / headword]
\ps·	[part of speech]
\pn·	[\ps for national language]
\ge·	[gloss—English]
\re·	[reversal—English]
\de·	[definition—English]
\gn·	[gloss—national language]
\dn·	[definition—national language]
\rf·	[reference]
\xv·	[example—vernacular]
\xe·	[example—English translation]
\xn·	[example—national lg. translation]
\ee·	[encyclopedic information—English]
\en·	[encyclopedic info.—national lg.]
\lf·	[lexical function (lexical network)]
\le·	[\lf gloss—English]
\ln·	[\lf gloss—national language]
\mr·	[morphology]
\bw·	[borrowed word]
\cf·	[confer/cross-reference]
\ce·	[\cf gloss—English]
\cn·	[\cf gloss—national language]
\sd·	[semantic domain]
\st·	[status of entry]
\so·	[source]
\dt·	[date entry last worked on][8]

Additional field markers for *expanded entries* can be added as needed. See §2.1.

[8]This can be set up within SHOEBOX to activate the DATESTAMP feature for automatically updating when the record was last worked on.

\lx ama	← [lexeme / headword]
\ps n \ge F \re father ; uncle (paternal) \de male of first ascending... \gn ayah ; bapak \lf Cpart = ina \le mother \ln ibu	← Field markers entered by database template.
\lf Spec = ama ebanat \le biological father	← Additional field markers inserted within entry.
\lf Spec = ama haat \le father's eldest brother	← Additional field markers inserted within entry.
↓ ↓	← Down arrow represents additional fields not indicated here. Other fields from original database template.
\sd Nkin \dt 28/Feb/84	

4.6 Root-oriented vs. lexeme-oriented databases

The compiler must decide early on whether to organize the dictionary around the root morphemes (structure-centric units) or around the surface-form lexemes (meaning-centric units). This is a significant issue, particularly with prefixing languages.

Landau (1989:33) notes:

> According to one scholar, the four basic systems of classification are by the alphabet, by the form of the entry words (morphemic), by meaning (semantic), or by no system at all (haphazard). The great advantage of the alphabet is that everybody knows it. A morphemic arrangement, which links words sharing a common form, such as *mishap* and *happen* or all the forms endings in *-ology*, would be of interest mainly to linguists. Semantic arrangements are employed in some thesauruses that, however, also have extensive alphabetical indexes to refer the reader to the various conceptual categories associated with each term.

The primary consideration here goes back to audience and purpose. Despite Landau's claim, it is not the case that "everybody knows" the alphabet.[9] Linguists tend to want to organize dictionaries around root morphemes for their own convenience. However, local audiences (and often others, including other scholars) generally find it difficult to find information organized around the root morphemes. They usually look for the surface form

[9]Many literacy programs for preliterate societies, non-formal education, or adult vernacular literacy, while teaching the letters of the alphabet, often fail to teach the alphabet as a conventionalized ordering of letters for mnemonic and organizational purposes. This then fails to equip the new readers with a basic skill needed to access tools, such as dictionaries, that build bridges for survival in the larger world.

first and then give up.[10] For example, in Buru they would want to look up **enyikut** under **en...** rather than under the root **iko**, and **ekhida** under **ek...** rather than under the root **hida**. Unfortunately, most major dictionaries of Austronesian languages have been heavily root-oriented.[11] Both strategies have their advantages and disadvantages, some of which are discussed below (see §4.6.1 for a summary). It is best to choose one strategy as primary over the other (root-oriented vs. lexeme-oriented, although a marriage of the two is possible) keeping in mind the associated advantages and disadvantages. To accomplish both requires some sophisticated tweaking of the database that is beyond the skill of the novice or even the average compiler.

OUR RECOMMENDATION: We recommend essentially a lexeme-based dictionary that also contains basic entries for root morphemes and affixes to show the morphological parts of the language and also to handle interlinearization.

Not every surface form should be in the lexicon. Some languages have classes of words, such as verbs, inflected for person and number with no other change in the meaning (e.g. *amo, amas, ama, amamos, aman*, or *'ala, mala, nala, tala*). For these types of words, only the *citation form* (discussed under **\lc** in §2.1, and in more detail in §5.4.4) should be an entry in the dictionary. The other forms should be derivable from information in the grammatical introduction to the dictionary. If there is an irregularity in the paradigm, that would be laid out overtly using the appropriate person-number form of the paradigm fields.

The two database formats (root-based vs. lexeme-based) might look something like the following:

Root-based DB (structure)

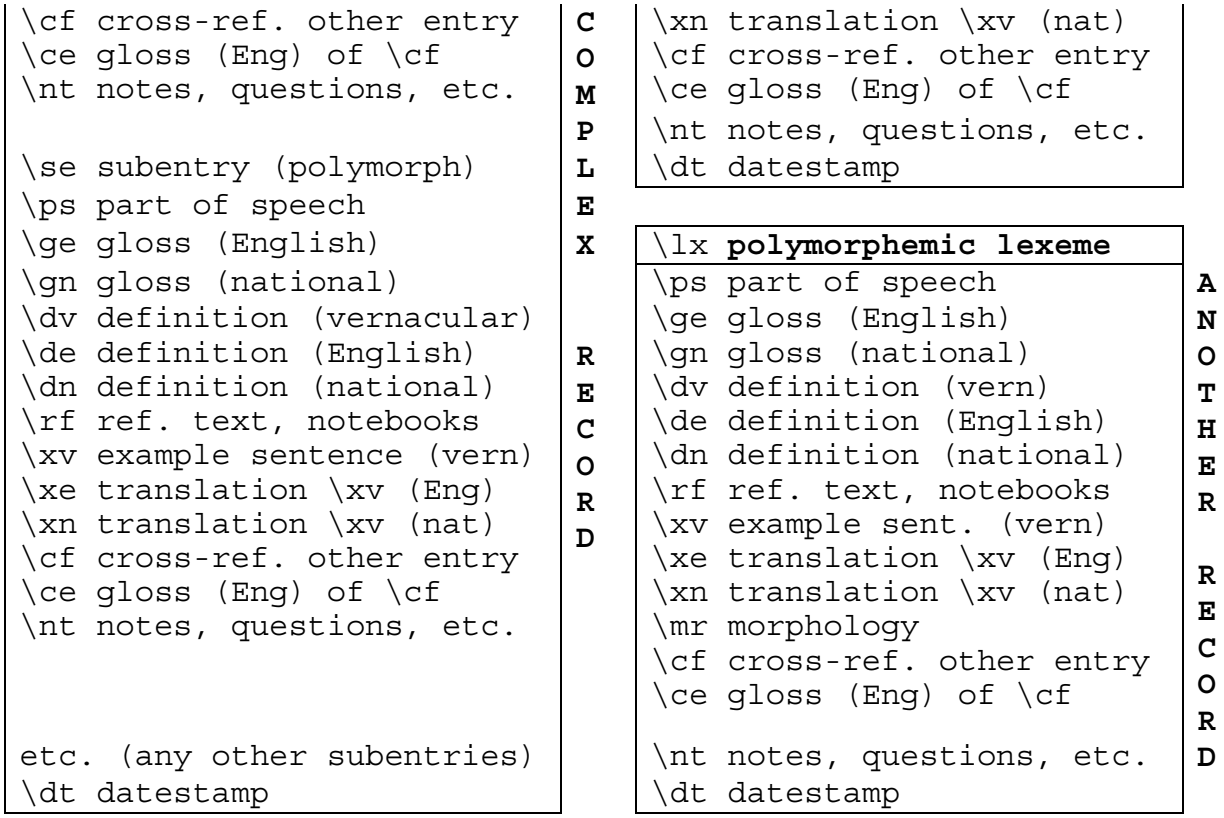
\lx **root** lexeme	
\ps part of speech	**J**
\ge gloss (English)	**U**
\gn gloss (national)	**S**
\dv definition (vernacular)	**T**
\de definition (English)	
\dn definition (national)	
\rf ref. text, notebooks	**O**
\xv example sentence (vern)	**N**
\xe translation \xv (Eng)	**E**
\xn translation \xv (nat)	

Lexeme-based DB (meaning)

\lx **root** lexeme	
\va list of variants	**O**
\ps part of speech	**N**
\ge gloss (English)	**E**
\gn gloss (national)	
\dv definition (vern)	**R**
\de definition (English)	**E**
\dn definition (national)	**C**
\rf ref. text, notebooks	**O**
\xv example sent. (vern)	**R**
\xe translation \xv (Eng)	**D**

[10]It can take a major effort to educate a whole society to parse words to find the root morphemes, and the organizational infrastructure required to do so may not exist. By contrast, many people who know how to read national languages learned the order of the alphabet in the process of learning to read, whether or not they attended a formal school.

[11]Zorc (1992) gives a negative critique of the heavily root-oriented Austronesian dictionaries pointing out that experience with end-users favors a combined approach.



In the root-based database, polymorphemic lexemes related to the root are seconded under the root form and become a part of the entry for the root form—this approach is structure-oriented. In the lexeme-oriented database, each lexeme has its own entry and the relationship that exists between root lexemes and polymorphemic lexemes based on that root are handled by cross-referencing using the **\lf, \cf, \va,** and **\mn** bundles of fields, just as headwords that may not be based on that root are handled—this approach is meaning oriented.

The lexicographer biased in favor of a root-based (form) approach might organize *hairbrush, toothbrush,* and *paintbrush* under the headword *brush*. The lexicographer biased in favor of a lexeme-based approach would argue that languages are full of lexemes such as *remove* which is clearly not synchronically the sum of *move* plus *re-* and must be handled in terms of meaning, not form.

Root-based approach

```
\lx brush
\ps n
\ge bristly_intrument
\de bristly instrument used
    for cleaning, arranging, or
    applying a liquid to s.t
\se hairbrush
\ps n
\de k.o. brush typically with
    stiff one inch long
    bristles loosely spaced
    arranged perpendicularly to
    the handle for rearranging
    hair
\se toothbrush
\ps n
\de k.o. brush with stiff one-
    quarter inch bristles
    tightly spaced arranged
    perpendicularly to the
    handle for cleaning teeth
\se paintbrush
\ps n
\de k.o. brush of varying sizes
    and varying lengths and
    textures of bristles
    arranged as an extension of
    the handle used to apply
    paint and similar materials
```

brush *n.* bristly instrument used for cleaning, arranging, or applying a liquid to s.t.

hairbrush *n.* k.o. brush typically with stiff one inch long bristles loosely spaced arranged perpendicularly to the handle for rearranging hair.

toothbrush *n.* k.o. brush with stiff one-quarter inch bristles tightly spaced arranged perpendicularly to the handle for cleaning teeth.

paintbrush *n.* k.o. brush of varying sizes and varying lengths and textures of bristles arranged as an extension of the handle used to apply paint and similar materials.

The lexicographer biased in favor of a lexeme-based (meaning) approach, would argue that *hairbrush, toothbrush,* and *paintbrush* are types under the generic *brush* and are unique lexemes in the language, part of the conventionalized knowledge bank of the culture, each with its own associated activities, materials, and industries, and should be handled as follows:

Lexeme-based approach

```
\lx brush
\ps n
\ge bristly_instrument
\de bristly instrument used for
    cleaning, arranging, or
    applying a liquid to s.t
\lf Part = handle
\le ...
\lf Part = bristles
\le ...
\lf Spec = hairbrush
```

brush *n.* bristly instrument used for cleaning, arranging, or applying a liquid to s.t. *Part:* **handle** '...'; *Part:* **bristles** '...'; *Spec:* **hairbrush** '...'; *Spec:* **toothbrush** '...'; *Spec:* **paintbrush** '...'; *Spec:* **mustache brush** '...'.
— *v.* to use a brush (n).

```
\le ...
\lf Spec = toothbrush
\le ...
\lf Spec = paintbrush
\le ...
\lf Spec = mustache brush
\le ...
\ps v
\de to use a brush (n)
```

```
\lx hairbrush
\ps n
\de k.o. brush typically with
    stiff one inch long
    bristles loosely spaced
    arranged perpendicularly to
    the handle for rearranging
    hair
\lf Gen = brush
\le ...
```

hairbrush *n.* k.o. brush typically with stiff one inch long bristles loosely spaced arranged perpendicularly to the handle for rearranging hair. *Gen:* **brush** '...'.

[*One approach — use* **\lf**]

```
\lx hairbrush
\ps n
\de k.o. brush typically with
    stiff one inch long
    bristles loosely spaced
    arranged perpendicularly to
    the handle for rearranging
    hair
\cf brush
\ce ...
```

hairbrush *n.* k.o. brush typically with stiff one inch long bristles loosely spaced arranged perpendicularly to the handle for rearranging hair. *See:* **brush** '...'.

[*Another approach — use* **\cf**]

A *root-based* database is keyed to root morphemes (and also includes bound morphemes like **-ku**). A root-based approach is often favored for morpheme-level analysis for interlinearizing texts. Generally there are no polymorphemic words found in any key field. Rather, these polymorphemic forms and their related information would be found under the related root form.

Root-based approach (structure oriented)

```
\lx bersih
\ps adj
\ge clean
\se kebersihan
\ps n
\ge cleanliness
\se membersihkan
\ps v
\ge to clean
```

← Beginning of record
← 1st subentry
← 2nd subentry
← 3rd subentry

In a *lexeme-based* database, on the other hand, the above subentries would be organized separately as full lexical entries that are cross-referenced back to **bersih**. Such a lexeme-based approach is preferred by many lexicographers because it focuses on the 'meaning chunks' irrespective of the root. These separate lexical entries are then cross-referenced back to their root form through the **\lf, \cf \mn,** or **\mr** field bundles. The separate but related lexical entries can be created and filled in from within the root entry through the use of SHOEBOX's JUMP feature (ALT + F6).

Lexeme-based approach (meaning oriented)

```
\lx bersih
\ps adj
\ge clean
\cf kebersihan
\ce cleanliness
\cf membersihkan
\ce clean s.t.
```

bersih *adj.* clean. *See:* **kebersihan** 'cleanliness'; **membersihkan** 'clean s.t.'.

```
\lx kebersihan
\ps n
\ge cleanliness
\mr ke-bersih-an
\cf bersih
\ce clean
```

kebersihan *n.* cleanliness. *Morph:* **ke-bersih-an**. *See:* **bersih** 'clean'.

```
\lx membersihkan
\ps vt
\ge clean
\de clean s.t
\mr meN-bersih-kan
\cf bersih
\ce clean (adj.)
```

membersihkan *vt.* clean s.t. *Morph:* **meN-bersih-kan**. *See:* **bersih** 'clean (adj.)'.

For sanity's sake it is important to also cross-list these polymorphemic entries in the root entry (e.g. using **\cf** or **\lf**). Otherwise the compiler would soon forget which related forms had already been addressed in the lexicon (because, being separate entries, they would be sorted alphabetically into their appropriate places).

Alternatively, the entry for the root **bersih** above could be more specific in the relationship between the forms by using the **\lf** fields rather than the **\cf** fields (see chapter 7).

```
\lx bersih
\ps adj
\ge clean
\lf Nres = kebersihan
\le cleanliness
\lf Cause = membersihkan
\le clean s.t.
```

bersih *adj.* clean. *Nres:* **kebersihan** 'cleanliness'; *Cause:* **membersihkan** 'clean s.t.'.

4.6.1 Comparing the two approaches

Root-Based Format	Lexeme-Based Format
a. full root-related network in one entry	a. root-related network indexed to other entries
b. morpheme-level interlinearization of texts	b. word- and morpheme-level interlinearization
c. many complex entries	c. complexity in cross-referencing other entries
d. polymorphs often underrepresented	d. quick updating of polymorphs from texts
e. tends to be frustrating to average user	e. frustrating to linguists looking for morphological unity
f. structurally driven	f. semantically driven

4.6.2 Advantages and disadvantages

For data management purposes, the main advantage to the lexeme-based format is that one has instant access to the polymorphemic forms (since SHOEBOX will index all **\lx** fields). One can relatively easily confirm that all of the principal or significant polymorphemic lexemes have been accounted for by comparing the sorted lexeme-based database with a comprehensive word list of all one's texts. This would be nearly impossible under the root-based approach.[12] Also, if one is principally interested in building a dictionary and annotated texts built around *words* (not morphemes), then this lexeme-based format is again the one of choice.

A relevant consideration here is that one can easily extract a lexeme-based database from a root-based database (by converting all **\se** codes to **\lx** codes, with some post-editing), whereas developing a subentry structure from a lexeme-based database is far more complex (requiring tight consistency in the use of **\cf** fields, for example) and would require far more post-editing. There are also problems with sorting polymorphemic forms into the correct homonym and with the ordering of the resulting subentries within their main root entry.

The lexeme-based approach *requires* careful inclusion of forward and back cross-references between related lexemes (using **\lf** bundles**, \cf** bundles, and **\mr, \mn,** and **\va** bundles). If this cross-referencing is forgotten, there will be detached entries floating around which cannot be related to their roots and other morphologically related entries.

There are political considerations here as well. Irrespective of the compiler's preference or the local community's ability to use the final product, there may be strong pressures or regulations from some institution such as a national language institute or the department

[12]But this could be done *outside* the original database. It would simply require the database to be copied to another file, and the **\se** codes converted to **\lx**. The new database could be read into SHOEBOX and resorted. If one were to try to use this resorted database, it would be with the understanding that some fields relevant to the original root morpheme (the first **\lx**) are probably now repackaged as part of the last subentry (**\se**). Version 2.0 of SHOEBOX can compare the resulting **\lx** contents against a text corpus using the SPELL CHECKER feature.

of education in a country requiring conformity to one sort of organization over another. There may be traditions for how dictionaries in a region or in a language family are organized. These issues should be investigated early in the development of a dictionary.

4.6.3 A suggested compromise

There is nothing (except perhaps government regulations) that requires an either/or approach. A satisfactory solution is a marriage between the two approaches. Lexemes, whether monomorphemic or polymorphemic, can be organized as individual headwords (**\lx**). Roots and other morphemes can also be entered as individual headwords (**\lx**). The lexical database can thus serve as both a structural base for interlinearizing texts and a meaning base for organizing the cultural-linguistic units of the language. In this approach the burden is on the compiler to be ruthless in cross-referencing (using **\lf** bundles**, \cf** bundles, and **\mr, \mn,** and **\va** bundles). This compromise incorporates the advantages of both root-based and lexeme-based approaches, and solves some of the disadvantages associated with either approach by itself.

Examples from Tetun (West Timor, Indonesia) show how information can be organized in this fashion. The series of entries that follow are inter-related and demonstrate how roots, other morphemes, and polymorphemic forms can be handled in the compromise approach we recommend.

```
\lx ai
\ps n
\sd Nplant
\sn 1
\ge tree
\sn 2
\ge wood
\lf Nres = ai balun
\le casket
\lf Nres = ai kabelak
\le board
\et *kaSiw
\eg wood
```

ai *n.* 1) tree. 2) wood. *Nres:* **ai balun** ‘casket’; *Nres:* **ai kabelak** ‘board’. *Etym:* *kaSiw ‘wood’.

[monomorphemic root; cross-referencing polymorphemic forms]

```
\lx ai balun
\ps n
\sd Ncult
\ge casket ; coffin
\mr ai balu-n
\cf ai
\ce wood
\cf balu
\ce side, part
```

ai balun *n.* casket, coffin. *Morph:* **ai balu-n**. *See:* **ai** ‘wood’; **balu** ‘side, part’.

[polymorphemic lexeme; identifying **ai, balu,** and **-n**]

```
\lx ai kabelak
\ps n
\sd Ncult
\ge board ; plank
\mr ai ka-bela-k
\cf ai
\ce wood
\cf bela
\ce flat
```

ai kabelak *n.* board, plank. *Morph:* **ai ka-bela-k**. *See:* **ai** 'wood'; **bela** 'flat'.

[polymorphemic lexeme; identifying **ai, bela, ka-,** and **-k**]

```
\lx balu
\ps n
\ge part ; side ; half
\lf Spec = mota balu
\le (other) side of river
\lf Spec = balu-balu...,
    balu-balu...
\le half of (group)..., the
    other half...
\cf balun
\ce side
```

balu *n.* part, side, half. *Spec:* **mota balu** '(other) side of river'; *Spec:* **balu-balu..., balu-balu...** 'half of (group)..., the other half...'. *See:* **balun** 'side'.

[monomorphemic root]

```
\lx balun
\ps n
\ge side ; remainder ; some
\lf Idiom = ai balun
\le casket (lit. 'its
    wooden sides')
\mr balu-n
\cf balu
\ce part
```

balun *n.* side, remainder, some. *Idiom:* **ai balun** 'casket (lit. 'its wooden sides')'. *Morph:* **balu-n**. *See:* **balu** 'part'.

[polymorphemic lexeme; identifying **balu** and **-n**]

```
\lx bela
\ps vn
\ge flat ; level
\cf kabelak
\ce flat (adj)
\cf belak
\ce flat round chest disk
\cf kabelan
\ce side, face
\cf belar
\ce spread out, multiply
```

bela *vn.* flat, level. *See:* **kabelak** 'flat (adj)'; **belak** 'flat round chest disk'; **kabelan** 'side, face'; **belar** 'spread out, multiply'.

[monomorphemic root; identifying related polymorphemic forms]

```
\lx ka-
\hm 2
\ps Vpref
\ge STAT
\re stative ; be
\de be; stative prefix
    deriving adjectivals
    from non-active verbs
\va k-
```

ka-$_2$ *Vpref.* be; stative prefix deriving adjectivals from non-active verbs. *Variant:* **k-**.

[general prefix; information required for morpheme-level interlinearizing]

```
\lx -k
\hm 2
\ps Nsuf
\ge NOM
\re *
\de nominal suffix indicating
    an independent unit (in
    contrast with the part-
    whole relationship
    expressed by the genitive
    fv:-n)
```

-k$_2$ *Nsuf.* nominal suffix indicating an independent unit (in contrast with the part-whole relationship expressed by the genitive **-n**).

[general suffix; information required for morpheme-level interlinearizing; no reversal]

```
\lx -n
\ps Nsuf
\ge GEN
\re *
\de genitive suffix normally
    indicating a part-whole
    relationship
```

-n *Nsuf.* genitive suffix normally indicating a part-whole relationship.

[general suffix; information required for morpheme-level interlinearizing; no reversal]

Under this combined strategy bound roots do not necessarily require a citation form. Two alternatives for handling bound roots are presented below. The decision between the two approaches is left to the compiler's preference.

```
\lx bani-
\ps Rt
\ge F-in-law
\re *
\de father-in-law
\mn banin
```

[*Approach 1*]
bani- *Rt.* father-in-law. *See main entry:* **banin**.

[two entries; no reversal on root; use **\mn**; these **\ps Rt** entries can be 1) in the main lexicon, 2) in a separate database, or 3) can be removed from the main lexicon before processing in MDF as desired]

```
\lx banin
\ps n
\sd Nkin
\ge F_in_law
\re father-in-law
\de father-in-law
\lf Cpart = ki'i
\le mother-in-law,
    father's sister
\lf Idiom = ai fehuk banin
\le rotten cassava
\mr bani-n
```

banin *n.* father-in-law. *Cpart:* **ki'i** 'mother-in-law, father's sister'; *Idiom:* **ai fehuk banin** 'rotten cassava'. *Morph:* **bani-n**.

[polymorphemic form; **\mr** can be used by SHOEBOX INTERLINEAR function]

```
\lx bani-
\lc banin
\ps n
\sd Nkin
\ge F_in_law
\re father-in-law
\de father-in-law
\lf Cpart = ki'i
\le mother-in-law,
    father's sister
\lf Idiom = ai fehuk banin
\le rotten cassava
\mr bani-n
```

[*Approach 2*]

banin *n.* father-in-law. *Cpart:* **ki'i** 'mother-in-law, father's sister'; *Idiom:* **ai fehuk banin** 'rotten cassava'. *Morph:* **bani-n**.

[single entry using **\lc**; SHOEBOX INTERLINEAR function will see only **bani-**, so a separate parse database or use of semiautomated parsing is required for handling polymorphemic forms like **banin**; **\mr** field here is for printing purposes only, not for interlinearizing]

The first approach above incorporates information about morpheme breaks into the main lexicon for interlinearizing, whereas the second approach uses a separate PARSE.DB as a place for SHOEBOX to look for directions about parsing polymorphemic words into their underlying morphemes. The first approach, if the root morpheme entries are included in the main lexicon, will have a certain amount of redundancy. However, not all languages have simple, single, or predictable forms that are built from the root, so the first approach would be entirely appropriate. The second approach requires the compiler to anticipate the final printing view to keep everything ordered properly.

5. Structuring the database

5.1 Using a database structure vs. using unstructured text files in a word processor

Many compilers of dictionaries use only a word processing program to enter the data in what they expect to be the final form.

> **utan** *n*. non-bulbous edible leafy and stalky plant and fungi; including vegetables and mushrooms. *Spec:* **uta lafut . . .**

Word processors as a tool have many *disadvantages* for compiling a dictionary, only some of which can be compensated for using a stylesheet or document template. For example, sorting (alphabetizing) is often done manually, particularly with non-default sequences (e.g. sorting digraph **ng** separately after **n**; **ch** separately after **c**, etc.). Searching or jumping to nonadjacent entries is slow and cumbersome on large lexicons, even with 'fast' computers and hard disks. Reversing the dictionary (e.g. vegetable *n* **utan**; mushroom *n* **utan**) must be done manually with great tedium and a tremendous waste of time. Editorial changes (e.g. the publisher insists on headwords being all caps or underlined, or on part of speech being non-italic caps) or font changes required by switching to a different printer must often be done manually, entry-by-entry. Styles can be forgotten or flags misspelled when they are applied manually (e.g. *See:* occasionally not italicized or no colon or misspelled). Additional language information (such as the national language) would either clutter the entry visually or have to be handled separately with a reduplication of effort. Extracting subsets of information for analysis or separate publication (e.g. selecting out entries related to kinship and social relations, or plant terms) is extremely difficult. Housekeeping information (e.g. date last worked on, source of information, reference to notebook or text, etc.) is left out altogether, hidden, or deleted manually prior to publication. The disadvantages go on and on.

A lexicon well structured as a database overcomes these problems, particularly when using a computer program like SHOEBOX and piping the output through a utility like MDF to make the print format, labels and styles automatic and consistent. The focus in compiling the dictionary is then on structuring the lexical information rather than on formatting. The disadvantage is that one cannot see the final formatting until the database is run through a program like MDF. An entry like the one above might be entered as:

```
\lx utan
\ps n
\sd Nplant
\ge veg
\gn sayur ; jamu
\re vegetable ; mushroom
\de non-bulbous edible leafy
    and stalky plant and fungi
```

utan *n*. non-bulbous edible leafy and stalky plant and fungi.

With a database structure, information not relevant to a particular audience or purpose (in this case national language information) is ignored; formatting is automated (**\lx** converts to style and point size defined for headword, **\ps** for part of speech, **\cf** can be replaced consistently by italics *See:*, etc.). Fields such as **\sd** can be used for extraction and retrieval of plant terms (using SHOEBOX filters), **\ge** can be selected by the computer for a cursory interlinear gloss, while the words in **\re** can be used for the English finderlist automatically creating entries under both *vegetable* and *mushroom*.

TIP: The compiler should use the codes and format recommended in this *Guide*, whether the lexical database is compiled on paper by hand, in a word processor, or directly in SHOEBOX. This not only provides more possibilities to the compiler if they do decide eventually to make it a computerized database, but will also facilitate reversal of the dictionary, and recovery of the information if the data eventually needs to be processed by someone else posthumously.

5.2 Multiple language information (bilingual/multilingual lexical databases)

If several types of information are to be kept in more than one language (e.g. vernacular, international language-English, national language, regional language), MDF provides a consistent system to assist with this:

\g*v*	gloss *vernacular*
\g*e*	gloss *English*
\g*n*	gloss *national* language (Indonesian, Filipino, Thai, Spanish, French, Portuguese, Tok Pisin)
\g*r*	gloss *regional* language (Ambonese Malay, Kupang Malay, Ternate Malay, Manado Malay, Makasar Malay, Jakarta Malay, Cebuano, Swahili, etc.)

Reversal codes are used where what is required for interlinearizing is less than or different from the gloss fields.[1] See §2.3.

\re	reverse English
\rn	reverse national language
\rr	reverse regional language

Additional multilingual bundles of field markers are used:

[1]Many people interlinearize only in English, with a few also using the national language. Unless one foresees interlinearizing in more than one language it is not economical to use two full sets of gloss and reversal fields.

definition/description	example sentence[2]	word-level gloss	cross-reference
	\rf [reference]		
\dv	\xv [see §6.2]		\cf
\de	\xe	\we	\ce
\dn	\xn	\wn	\cn
\dr	\xr	\wr	\cr

usage	lexical functions	restrictions (only)	encyclopedic
\uv	\lf [see §7]	\ov	\ev
\ue	\le	\oe	\ee
\un	\ln	\on	\en
\ur	\lr	\or	\er

variants
\va
\ve
\vn
\vr

For publication it is recommended that information relevant to different target languages (e.g. \ge and \gn) be printed separately, either in separate sections of the same publication or in separate publications. Keeping more than two languages together tends to be visually cluttering and makes dictionaries difficult for the average user. For a working draft, printing in triglot may be workable for some compilers and this option is available in MDF. See §4.3.

Not recommended for publication:

\lx ama
\ps n \pn kb \ge F \re father ; uncle (paternal) \de father, uncle (paternal) \gn ayah ; bapak ; paman

ama *n.* father, uncle (paternal), *ayah, bapak, paman...*

Recommended: different sections or separate publications:

ama *n*. father, uncle (paternal)...

ama *kb*. ayah, bapak, paman...

The system incorporated here (v=vernacular, e=English, n=national language, r=regional language) should be flexible enough to handle the majority of situations. In many situations the regional language set of codes is not needed for the local situation and can

[2]An additional field that relates to this bundle is the **\xg** field, if the example sentence is to be interlinearized. This is not currently supported in MDF.

be used for a second national language. In the current configuration of MDF the regional language codes are tied to print when the national language options are selected. *They do not function independently* so they should not be used for other categories of language such as the researcher's national language like Finnish, Italian, Korean, or French.

5.3 Categories of information in a lexical entry

Ignoring formatting purposes for the moment, there are basically three general categories of information in a lexical entry: 1) information about the headword, 2) information about words related to the headword, and 3) housekeeping information.

5.3.1 Information about the headword

Most field markers in a record relate directly to the headword. These include: [NOTE \xx+ indicates a bundle of related fields.]

\lx	lexeme, lemma, headword	
\ph	phonetic [if not transparent from orthography]	
\sn	sense number	
\ps	part of speech	
\ge+	gloss	(\gv, \gn, \gr)
\re+	reversal	(\rn, \rr)
\de+	definition	(\dv, \dn, \dr)
\xv+	example sentence	(\xe, \xn, \xr, \xg)
\ue+	usage	(\uv, \un, \ur)
\oe+	restrictions	(\ov, \on, \or)
\ee+	encyclopedic information	(\ev, \en, \er)
\mr	morphology	

5.3.2 Information about words related to the headword

Some field markers relate a headword to other entries or to additional information, thus tying it in with its lexical network. These include:

\hm	homonym number	
\lf+	lexical functions	(\le, \ln, \lr)[3]
\sy	synonym	
\an	antonym	
\nt+	notes	(\na, \nd, \ng, \np, \nq, \ns)
\pd+	paradigm [structural pattern or completeness]	(various)
\et+	etymology, historical	(\eg, \es, \ec)
\bw	borrowed word; loan source	
\cf+	cross-reference	(\ce, \cn, \cr)

[3]These are described and illustrated in chapter 7.

\sd	semantic domain	
\va+	variant forms	(\ve, \vn, \vr)
\mn	main entry form	

TIP: The JUMP feature in SHOEBOX <ALT+F6> allows the user to check the converse of information relating to other headwords, or to create new entries while within a record. This is a very powerful feature of SHOEBOX and should be mastered early.

5.3.3 Housekeeping information

Additional fields help keep track of the history and reliability of the information. Some of this information does not need to be published.

\rf	reference [to notebook or text—usually combined with \xv]
\bb	bibliographical reference [reference to publication expanding on headword]
\pc	picture [reference or graphics insertion for publication]
\so	source [name of native speaker]
\st	status [processed, check text, don't print record, etc.]
\dt	date last worked on [use DATESTAMP]

5.4 Sort sequences (alphabetizing)

Four issues come into play regarding sort sequences and MDF: 1) getting the secondary sort order of homonyms corrected, and understanding the consequences of doing so, 2) restoring customized primary sort sequences, 3) choosing whether to sort by the citation form (**\lc**) or by the head lexeme (**\lx**) in entries where both occur, and 4) sorting bound roots (e.g. **–edo**) in a consistent pattern.

5.4.1 Getting homonyms in the correct order

Many languages may require a sort sequence that is different from a simple **a b c d e...** For example, they may need **e é, n ñ, n ng** [digraph], **m mb n nd ng ngg**, or other sequences not easily handled by commercial software. Furthermore, the sorting should be automatic in that a new record should be automatically placed in the correct sort position without any effort by the compiler. This is easily handled by a program such as SHOEBOX dedicated to lexicography, adapting the SORT sequence in the GLOBALS menu. (MDF currently overrides this for printing, but custom sort sequences can be reinstated, as discussed in §5.4.2.)

In addition to the primary sort order, there are secondary considerations, such as the order of homonyms. Where the compiler has entered **\hm 1**, **\hm 2** into the database correctly, MDF ensures that homonyms are sorted correctly. SHOEBOX (through version 2.0) does not account for secondary sort sequences, and so homonyms can be reordered in relation

to each other each time one of them is edited in SHOEBOX. Note that different homonyms are structured as separate entries.

```
\lx baa
\hm 1
\ps AUX
\ge only
```

baa$_1$ *AUX.* only.

```
\lx baa
\hm 2
\ps n
\ge stem
```

baa$_2$ *n.* stem.

There are a wide range of options in published dictionaries for indicating homonyms. MDF uses the subscript (e.g. **baa**$_1$, **baa**$_2$) as one that is common, visually pleasing, and easy to implement consistently on the computer. MDF provides for numbers in vernacular fields to automatically subscript, assuming that they cross-reference a particular homonym.[4]

```
\lx rahek
\ps AUX
\ge only
\lf Syn = baa1
\le only
```

rahek *AUX.* only. *Syn:* **baa**$_1$ 'only'.

5.4.2 Restoring customized primary sort sequences

Because MDF resorts the database to order homonyms correctly, this means it ignores any custom sort sequences set up in SHOEBOX.[5] The following steps maintain or restore a customized sort order:[6]

1) Copy the file MDFDICT.ANS to MDFANS.SAV.

---

[4]For those who need superscripted tone numbers within vernacular fields, we suggest marking the tones with otherwise unused symbols in SHOEBOX and then post-edit the MDF output in WORD, replacing those symbols with the appropriate superscripted numbers.

[5]To get primary (**\lx**) and secondary (**\hm**) fields both involved in the sorting, MDF uses the SIL program SRT, which uses a different command structure for defining the sort order than that used by SHOEBOX. Thus it was not possible to have MDF find and read the SHOEBOX sort command sequence and incorporate it for SRT.

[6]Compilers working on dictionaries in Spanish-speaking countries should be aware that the 10th Annual Congress of the Association of Spanish Language Academies voted in April 1994 to eliminate 'ch' and 'll' from the Spanish alphabet. Words beginning with these letters will now be listed under 'c' and 'l' respectively (reported in the *Charlotte Observer*, 30 April 1994). We are intrigued, since this move is probably driven by the inconvenience or inability of many commercial computer programs to perform non-ASCII or digraph sorts—sorts which are handled easily by SHOEBOX and MDF. Dictionary compilers should check that the country in which they work subscribes to these proposed changes before incorporating them by restructuring their lexicon (through a new sort order).

2) Edit MDFDICT.ANS with a text editor that can save the file as 'Text only' or 'ASCII'. *Do not make any other changes than those noted here!*

3) In MDFDICT.ANS insert the changes in the **\m** field. If, for example, one wishes to sort the digraphs **nd**, **ng**, the trigraph **ngg** and the monograph **ñ** separate from and following the **n**'s, then the following change would be made:

```
\m @ a b c d e f g h i j k l m n o p q r s t u v w x y z
   0 1 2 3 4 5 6 7 8 9 { | } ~ ! " # $ % & ' ( ) * + , . /
   : ; < = > ? [ \ ] ^ _ `

\m @ a b c d e f g h i j k l m n ñ nd ng ngg o p q r s t u v
   w x y z 0 1 2 3 4 5 6 7 8 9 { | } ~ ! " # $ % & ' ( ) * +
   , . / : ; < = > ? [ \ ] ^ _ `
```

4) Save the file as 'Text only' or 'ASCII' as MDFDICT.ANS and then test MDF on a sample database that includes data that should be effected by the changes (such as headwords that begin with **ñ** and **ng**).

5) If it works correctly, then process the entire database through MDF.

6) Once that is done, then post-edit the file in WORD *copying* another section header (the letters and line that appear before each new letter in the alphabet) to the correct place and adjusting it to reflect the changes. This is most easily done if *non-printing characters* are visible on the screen (set through the OPTIONS menu in WORD). Be sure to copy the appropriate division breaks and paragraph marks as well.

7) If MDF on your computer will be used by other users or for other languages, remember to copy MDFANS.SAV back to MDFDICT.ANS when you are done.

5.4.3 Sorting bound morphemes

An additional sort consideration is where bound roots with preceding hyphens, or suffixes are sorted. Our experience suggests that uninitiated dictionary users find the form better with the hyphenated form *following*, rather than preceding similar forms (e.g. **eta**$_1$, **eta**$_2$, **-eta**; rather than **-eta, eta**$_1$**, eta**$_2$). This needs to be tested locally. To ensure the hyphened forms sort as desired in SHOEBOX, under the GLOBAL SORT menu check that the hyphen is placed in parentheses (-) at the end of the **\srt** fields. MDF orders bound morphemes at the end by considering the hyphen as a secondary sort character (**\s** - in the MDFDICT.ANS control file).

5.4.4 Sorting citation forms (\lc)

Where a language has lexemes in a variety of inflected forms, none of which is 'basic', a citation form must be listed as the headword. While Romance languages such as Spanish have verbal infinitive forms such as *salir* 'to leave', Attic Greek references developed the convention of first person singular citation forms such as *baptizo* 'I dunk (e.g. cloth for dyeing), I immerse s.t.'

Speakers of languages with no written tradition may also have preferences for which form is used. Speakers of related languages in the same region may have different preferences for the citation form. In the province of Maluku, Indonesia (with around 130 languages), some language communities prefer first singular, some third singular, some first plural, some third plural. The preference is not only evident after extensive fieldwork, but is often evident in responses given upon taking an initial wordlist.[7]

A single language may reflect different preferences for different parts of speech. For example, Buru speakers clearly prefer to cite verbs in the first person plural (e.g. **ma iko** 'we go'; **ma kaa** 'we eat'), whereas human body part terms are evenly divided in responses between third singular and first plural forms (e.g. **kadan/kadanan** 'leg'; **raman/ramanan** 'eye').

A further need for citation forms occurs where root morphemes are isolatable, but never occur by themselves as a surface form. This is the case with *precategorials* (see C. Grimes 1992, and §9.3.1.3 in this volume). For example, in Buru **mae–** never occurs by itself, but always with derivational morphology as **mae-n, mae-t,** or **mae-k**. A Buru person would never look for the root by itself, because the root is not a minimal word! Therefore a citation form is required.

MDF provides the option of sorting by the citation form (**\lc**) or by the headword (**\lx**) for entries that use **\lc**. The following entry can sort in the MDF-formatted dictionary under the B's (sort by **\lc**) or under the A's (sort by **\lx**). [The option of also printing the contents of the **\lx** field as in the example below is accomplished by answering 'yes' to the MDF-prompted questions "Do you want entries sorted by the citation form?" and "Do you want to include the \lx keyfield reference with the \lc field when it is formatted?" The default answer to this latter question is 'No', so the MDF user must explicitly choose 'Yes' for the file to print as below].

[7]Initial impressions must be corroborated by other evidence, since a wordlist-taking situation is often one in which miscommunication occurs.

\lx -ao
\lc bekeao \ps v \ge screech ; howl \ue Formal/ritual \lf SynR = bengeao \le common speech

bekeao (*from:* **-ao**) *v.* screech, howl. *Usage:* Formal/ritual. *SynR:* **bengeao** 'common speech form'.

If the choice is to sort the above entry under the A's, then the compiler may want to organize the data in the following way to make it visually obvious why the entry appears out of place:[8]

\lx -ao
\lc (beke)-ao \ps v \ge screech ; howl \ue Formal/ritual \lf SynR = bengeao \le common speech

(beke)-ao *v.* screech, howl. *Usage:* Formal/ritual. *SynR:* **bengeao** 'common speech form'.

[8]We recommend testing a wide cross sample of users to see which approach is preferred for a given community, and why.

6. Structuring information in lexical entries

6.1 Principles for choosing headwords

Most people have a fairly good intuitive sense of what a 'word' is, as it relates to the headword of an entry. Some clarification will help sharpen intuitions and enable principled decisions to be made about including and excluding 'words' from our dictionaries. The more different the language is typologically from those we are familiar with, the less confident we tend to be in our intuitions. A highly polysynthetic language (e.g. highlands Papua New Guinea) or a highly isolating language (e.g. mainland Southeast Asia) will be handled differently from languages such as English or Indonesian.

Simple monomorphemic words are fairly straightforward:

```
\lx fatu
\ps n
\ge rock
```

fatu *n.* rock.

```
\lx wae
\ps n
\ge water
```

wae *n.* water.

```
\lx iko
\ps vi
\ge go
```

iko *vi.* go.

Compounds that have phonological evidence of functioning as a unit are also fairly straightforward candidates for headwords.

```
\lx fathese
\ps n
\ge cliff
\lt rock-wall
\mr fatu-hese
```

fathese *n.* cliff. *Lit:* 'rock-wall'. *Morph:* **fatu-hese**.

```
\lx hektatak
\ps vt
\ge abandon_s.t
\lt flee-drop
\mr heka-tata-k
```

hektatak *vt.* abandon s.t. *Lit:* 'flee-drop'. *Morph:* **heka-tata-k**.

There are also combinations of words that do not show phonological changes, but clearly function in a language as distinct cultural concepts or units. They may be phrasal or even clausal. Often the combination of such units is different than the sum of its parts, indicating non-restrictive, conventionalized, or semantically bleached senses. They often indicate types of a kind. English 'words' of this sort include *blackboard* (often green),

hairbrush (very different in form and function from *paintbrush* or *toothbrush*), *Christmas, christen, in a rut*, *on the dole, up a creek, no ball* (cricket term).

```
\lx geba nega
\ps n
\ge adult
\lt person easy
```

geba nega *n.* adult. *Lit:* 'person easy'.

```
\lx ba sohik
\ps vt
\ge hope_for_s.t
\cf sohik
\ce wait
```

ba sohik *vt.* hope for s.t. *See:* **sohik** 'wait'.

```
\lx geba ka kaa geba
\ps n
\ge cannibal
\lt person who is characterized
    by eating people
```

geba ka kaa geba *n.* cannibal. *Lit:* 'person who is characterized by eating people'.

There is disagreement among some lexicographers as to whether these latter two types should be handled as subentries or as separate entries. If the primary audience is the local populace, the separate entry strategy is probably best, supplemented by cross-references. [See examples and discussion in §4.2 and §4.6].

These types of emic units, whether they are simple morphemes, compounds, phrasal, or clausal, are all good candidates for a 'headword'. Such structural variety is what drives lexicographers to use the term *lexeme*, rather than 'word' to describe these units.

Pawley (1993:30/3/93) describes two views of language that are in tension for compilers of dictionaries.

> Many of the ideas which people formulate in their language are highly subjective constructions, having only the most tenuous connections with objectively measurable things and events. Some of these subjective formulations may enter the linguistic tradition, becoming standardized ways of saying things. Thus, each language community develops a unique body of resources representing a particular worldview, a particular shared tradition which is part of its culture.
>
> In describing language as a device for encoding a particular culture, the object is not to achieve the most parsimonious specification of grammatical form-meaning pairings. *The object is to describe what it takes to use a language properly as a member of society. Part of this is knowing what things to say, when to say them and how to say them in conventional ways.* The culture encoding approach leads us to take a very different definition of the lexicon from the grammarian's. Instead of striving to keep the lexicon small we need to enrich it. In fact we apply the terms 'lexicon', 'lexeme' (or 'lexical item') and 'lexicalized' in ways quite different from the grammarian. *Now these terms are defined with respect to cultural facts as*

> *well as with respect to purely structural criteria*. Complex words and compounds, and perhaps phrases, are considered part of the speaker's cultural lexicon if we can show that they have entered the social tradition, that they have attained the status of social institutions, being recognized as conventional 'names of things', as 'terms' in a set or terminology, as 'set phrases', and perhaps as 'appropriate things to say'. All grammatical strings are not socially equal. We award special status to those strings that are culturally significant, even though they may also be perfectly grammatical. The upshot is an enormous increase in the number of lexemes compared to the ideal grammarian's dictionary. [emphasis added]

Pawley (1986) identifies a number of tests for English that may help determine whether or not something can be considered to be a lexicalized form, and thus a candidate for a *lexeme* (headword) in the above sense. Many of the tests are adaptable to other language situations as well. Some of the tests depend upon a written tradition. The following material is adapted from Pawley (1986) and most of the examples are also from that source:

1) ***The naming test***: Can the candidate for a lexeme be referred to in questions or statements such as the following: 'What is it called?' 'It is called X.' 'We call it X, but they call it Y.'

2) ***Membership in a terminological system***: This assumes a lexical network as discussed in chapter 7. Does X encompass other terms; can one say 'it (dog) is a kind of X (animal)' (=generic)? Is it a member of a set of similar things; can one say 'X (a chair) is a kind of Y (furniture)' (=specific)? Can it be used to show contrast; 'is it a kind of X (fruit), but not a Y (vegetable)'? Does it have synonyms or antonyms?

3) ***Customary status***: Does the use of the phrase imply certain behavior patterns, values, or sequences of activities that are known by society at large? They represent conventionalized knowledge. For example, expected behavior at the *front door* is different from at the *back door* (besides their participation in idioms), indicating that these function as cultural units (lexemes) that are more significant than the sum of the parts. Consider *go to the mosque, get off work, take a vacation*.

4) ***Legal status***: Some phrases have such status that they are codified in legal usage: *driving under the influence, breaking and entering, assault and battery*, *justifiable homicide*. Even so-called 'primitive' societies with unwritten languages have categories of this sort for dealing with things like marriage negotiations and litigations over land, property, and adultery.

5) ***Speech act formulas***: Every language has some formulas "which carry out conversational moves" (Pawley 1986:106). For example, *excuse me, how are you, ya'll have a nice day*, etc.

6) ***Use of acronyms***: This is often proof that a multi-word phrase represents concepts that have attained conventionalized or institutionalized status. Consider: *VIP, DWI/DUI, IQ, RBI, SAT, ASAP, PTO, PTL, AWOL, BS, RSVP, R and R*; in Indonesia: *KB, DKI, KK, ABRI, DPRD*, *GBHN,* etc.

7) ***Single-word synonyms***: *the only one of its kind ↔ unique*.

8) ***Belonging to a terminological set***: This is similar to (2), but focuses more on a pair of antonyms. Consider: *tell the truth ↔ tell a lie, take care of ↔ neglect*.

9) ***Base for inflected or derived forms***: *short-temper → short-tempered*; *ooh and ah → oohing and ahing,* Indonesian *ke mana → dikemanakannya* ('to where' → 'wind up where').

10) ***Internal pause unacceptable***: The unacceptability of inserting a pause in the middle of clichés, idioms, and compounds is partial indication of their functioning as a unit. Consider the functional differences between *bunch of baloney* vs. *bunch of bananas*. One can say *two bunches of bananas*, but cannot do the same with the figurative sense of *bunch of baloney*.

11) ***Inseparability of constituents***: Insertion of other material changes the unity or naturalness of a phrasal lexeme. Consider: *lead up the garden path.* Saying *lead up the beautiful garden path* shifts it from a figurative to a literal interpretation. This is similar to (10) above.

12) ***Ambiguity as to whether it should be written as a single word***: *whatchamacallit, thingamajig, man-in-the-street, oneupmanship.*

13) ***Conventionally reduced pronunciation***: *bosun* (boatswain), *won't, can't, o'clock, Newfoundland, Christmas, Worchestershire*, *thruppence* (three pence) etc.

14) ***Conventionally truncated forms***: Widespread occurrence of shortened forms often indicate their role as a lexeme in the language: *exam(ination), rad(ical), ex-con(vict), con(vict), con(fidence man), con(fidence trick), ex(-husband/-wife), pro and con,* etc.

15) ***Omission of headword***: The modifier stands metonymically for the whole: *She had an oral (examination), He had a physical (examination), A short (circuit) cut off the (electrical) power.*

16) ***Omission of final constituents***: This often implies conventionalized knowledge: *If you can't beat 'em..., A stitch in time..., I haven't the faintest (idea).* These elided forms are often marked by peculiar intonation.

17) ***Stress and intonation patterns***: Different languages give different phonological clues for what is seen to function as a unit. English often uses stress and intonation. Government jargon is often coined through these means. Consider *political matters memorandum* (see Pawley 1986:108).

18) ***Invariable constituents or grammatical frame***: The demanding and rhetorical *Who do you think you are*? does not have the same impact in the future. *Kick the bucket* does not mean the same when put in the passive. *The thought had crossed my mind,* and *he took the law into his own hands* are unnatural in the passive. Compare also stripped down formulaic sentences *easier said than done, spoken like a man*! There are also syntactically irregular or archaic idioms like *easy does it, no go, no way, be that as it may, (she) wants in, once upon a time.*

19) ***Use of definite article on first mention***: In English this can indicate the conventionalized nature of the 'object', showing the speaker assumes the identity is understood by the addressee: *the fire department, the foreign legion, the eight ball*.

20) ***Writing conventions***: Where there is a written tradition these may provide clues to perceived status as a unit. *Capitals* may indicate lexemes that are not typical proper nouns: *Third World, Big Bang, Inner City*. Beware that where a society has the luxury of supporting a literary community, some writers manipulate the use of capitals for unconventional purposes. *Quotation marks* may also indicate unitary status: *he was considered a 'bad boy'*. Orally, some speakers use *so-called* or a preceding pause to mark an equivalent to quote marks.

21) ***Unpredictability of form-meaning relation in semantic idioms***: *kick the bucket*, *chew the fat*, *shoot the breeze.*

22) ***Arbitrary selection of one meaning***: Notice that *button hole* is a hole FOR putting buttons THROUGH, whereas *bullet hole* is a hole MADE BY bullets, *posthole* is a hole FOR setting posts IN, etc.

23) ***Use in ritual language of parallelism***: This is a special case of (2) and (8). Ritual language in parallelisms is widespread. It is found, for example, in Biblical Hebrew and many Austronesian languages, particularly in eastern Indonesia (Fox 1988). Existence as a paired entity in this context is sufficient for justifying its status as a conventionalized unit, and hence a lexeme.

Refer to Pawley (1986) for additional examples and more detailed discussion.

6.1.1 Affixes

Affixes should be entered into the lexical database, both for the resulting dictionary and for interlinearizing. When entries for affixes are generated through the process of

interlinearizing, it is helpful to keep track of them on a piece of scrap paper and add the hyphen to the key field later, as appropriate. Entries for affixes tend to map grammatical functions and be less straightforward than entries for lexical roots.

```
\lx ep- [prefix]
\ps Vpref
\ge CAUS
\re causative
\de causative prefix, usually
    indicating direct causation
```

ep- *Vpref.* causative prefix, usually indicating direct causation.

```
\lx -n  [suffix]
\ps Nsuf
\ge 3sG
\re his ; hers ; its
\de his, hers, its; third
    singular genitive suffix,
    normally indicating a
    physical or conceptual
    part-whole relationship
```

-n *Nsuf.* his, hers, its; third singular genitive suffix, normally indicating a physical or conceptual part-whole relationship.

```
\lx <um>  [infix]
\ps Vinf
\ge UF
\re undergoer focus
\de undergoer focus marker
```

<um> *Vinf.* undergoer focus marker.

6.1.2 Lexical root plus affixes

Since dictionaries are normally organized on the principle of alphabetizing, *suffixing* languages do not tend to raise challenges for information organization and retrieval. The real challenge comes from *prefixing* languages. This returns us to the issue of audience [§4.2]. A scholarly audience may be able to handle bound roots (although perhaps not as easily as might be assumed). However, in many languages the bare bound root simply does not qualify as a minimal word or utterance, and so the local audience (=native speakers) would never look for it as the bare root. This is why a citation form is required (**\lc** see §5.4.4). One solution is as follows:

```
\lx -bate
\lc (ma)-bate
\ps n
\ge abundance
```

(ma)-bate *n.* abundance.

```
\lx -bafa
\lc (na)-bafa
\ps v
\ge ambush
\de wait in ambush
```

(na)-bafa *v.* wait in ambush.

MDF substitutes **\lc** for the headword when printing. This presents a dilemma, since until the local audience learns how to parse words (which takes an educational infrastructure and time) they may not know where to look up a word. MDF menu options allow the user to choose whether these entries should be sorted by the **\lx** field or the **\lc** field. Because of the nature of citation forms, sorting on the **\lc** will probably result in many languages in certain sections of the printed dictionary being disproportionally huge.

6.2 Choosing example sentences

Why are sentences like *See Spot run* or *Run, Spot, run!* not good example sentences for a lexicon?

An excellent discussion of example sentences is found in chapter 9 of Bartholomew and Schoenhals (1983). A few of their points are summarized here:

> Illustrative sentences serve both the compiler of the bilingual dictionary and its user. During the process of eliciting illustrative sentences, the compiler becomes aware of sense discrimination co-occurrence restrictions on classes of lexical items, or grammatical restrictions which he had overlooked. (1983:59)

They list as functions of example sentences:[1]

1) Delineate and exemplify sense discrimination.

2) Exemplify correct or unusual grammatical contexts.

3) Demonstrate legitimacy of glosses or translation equivalents.

4) Clarify potential ambiguities set up by the presence of multiple glosses.

5) Illustrate norms of local culture or local literary style.

In other words, a well-chosen example sentence can be made to work for you, highlighting some of the characteristics that may still be unclear from the definition.[2]

Good example sentences, of course, should be complete, grammatical, and preferably natural. In addition:

> A good illustrative sentence supplies a specific context which helps to define the word being illustrated. Such a sentence should include at least one of the salient characteristics of the word under consideration. in many instances it should be possible to deduce the meaning of the word even if one were unfamiliar with the

[1]These are partially rephrased for our purposes.

[2]This should not, however, become a substitute for the hard work of making good definitions.

gloss. Characteristic subjects or objects may be used with verbs to provide mental clues as to the specific action indicated. Other useful contextual ideas include instrument, location, or cause and effect relationship. (Bartholomew and Schoenhals 1983:60)

TIP: Many of the characteristic associations or typical co-occurrences should be mapped out in the lexical functions fields (**\lf** bundle—see chapter 7). Procedurally, we recommend not eliciting or selecting illustrative sentences for a lexeme until most of its lexical relations have been fully explored. Not only does this give the compiler a more rounded picture of what s/he is dealing with, but it also gives the language assistant(s) a broad and freshly explored context for thinking about example sentences. One can then concentrate on choosing example sentences that are *dynamic*, memorable, or even dramatic, as well as illustrative.

Bartholomew and Schoenhals (1983:61ff.) list with examples in Spanish and English the following "associational categories" which can be included in an illustrative sentence as context for the lexeme.[3]

1) ***Characteristic attribute***: He wore his *red* **berang** cloth across his chest to do the war dance. She used the *sharp* **katanan** to peel the cassava.

2) ***Characteristic behavior or action***: **Motin** causes recurring *fever, chills and shakes*. **Geba emsihi** often *stagger* home after drinking too much palmwine with their friends.

3) ***Characteristic use***: My father has a **kupan elen** *in which he keeps his valuables* out in the garden house. We use **kelambu** around our bed *to keep out mosquitos and other bugs.*

4) ***Characteristic position or location***: My father left his **waga** *at the shore* after paddling it *across the lake*. The warrior's **todo** is kept in its *scabbard*.

5) ***Characteristic material***: We used *split bamboo* to make our **hese** [wall] on our new house. Hunters make **suran** [spike traps] from *uka bolo* [bamboo sp.].

6) ***Characteristic subject, object, or instrument of an action***: When making a new garden we **fell** *the big trees with an axe*. The **enhero maen** [spear shaft] broke when the *wounded pig* dragged it through the underbrush.

7) ***Contrast, gradation, or complementary categories***: The **kori** represents the bride's interests in marriage negotiations, and the **sanat** the man's interests. The boy is

[3]We have adapted the examples to a Buru context.

emteno [heavy (of people)], but the gunny sack of copra is **beha** [heavy (of things)].

8) ***Cause-effect relationships***: He *drank palmwine* until he got **emsihi** and created a disturbance, because his elder sibling was not contributing to the bridewealth pool. The *cuscus rotted* because he forgot to **touk unet** [check his snares].

9) ***Abstractions or general classificatory terms***: When all the grain in the bin had been either eaten or planted, the *grainbin* was **fuun** [empty]. When he saw the *isaleu* [python] in the jungle, he felt **emgihi** [horrified and grossed out], and moved way quickly.

10) ***Part-whole relationships***: The *pig's* **ngisnap** [tusk] was four fingers long. The **sufen** [doorway] is where people go into the *house*.

11) ***Synonym or class name***: **Lian** [caves] are *holes* in cliffs big enough for people to sleep in. A **yoho** [civet cat] is small *animal* like a wild dog or cat that lives in the jungle...

12) ***Comparison***: **Gehut rali** doesn't have purple speckles like the *traditional taro* has. **Geb masi** [coastal people] do not know how to survive in the jungle as well as **geb fuka** [mountain people].

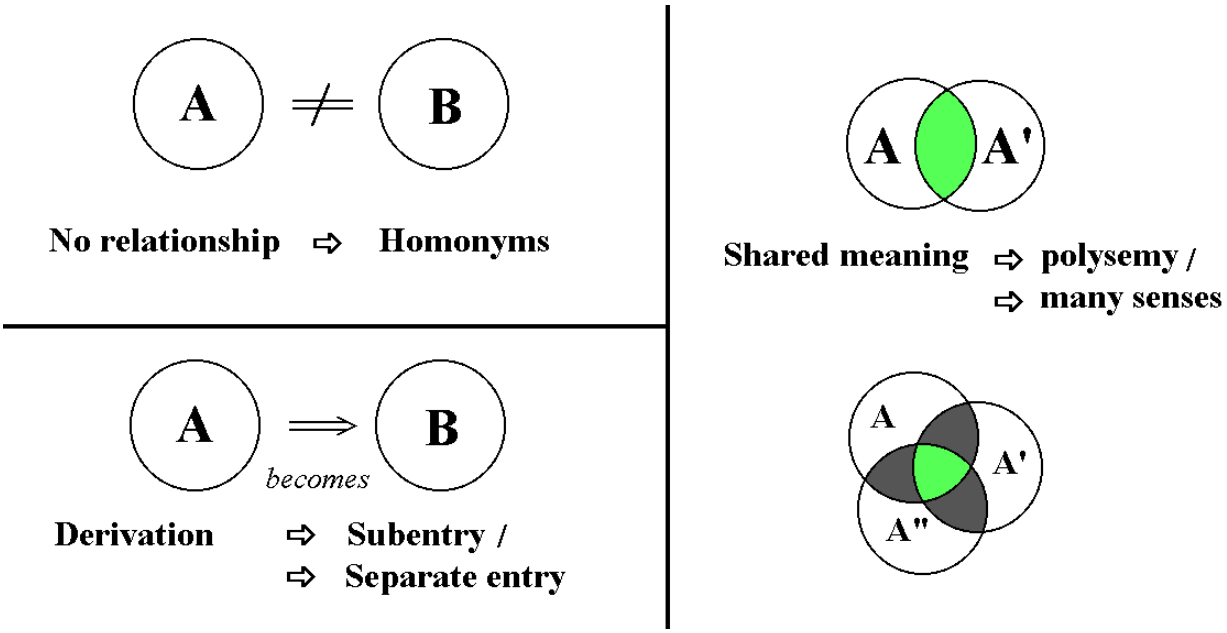
Bartholomew and Schoenhals (1983:64–69) also have a good discussion of do's and don'ts for obtaining good example sentences.

CAUTION: Avoid sentences created by non-native speakers or by the foreign compiler. And avoid using translated materials as a source for illustrative sentences. If one uses sentences extracted from natural text, remember that running text provides context. Extracting a sentence from that context often leaves it depending on implicit and presupposed information, or with anaphoric pointers that have nothing to point to. Thus, while the sentence makes perfectly good sense in context, it may seem incoherent or even ungrammatical to a native speaker when removed from context. It is thus important to edit and check such sentences with the assistance of a skilled native speaker before using them in isolation in the lexicon.

6.3 Different words or different senses? (homonymy vs. polysemy)

When a single form can function in more than one category without any explicit derivation, the lexicographer must decide whether to handle them as *homonymy* (same form but unrelated meaning, therefore separate lexemes), or as *polysemy* (same form with

range of related meanings, therefore subentries or multiple senses of the same lexeme).[4] The following figure illustrates various relationships between categories as they relate to homonymy and polysemy.



In one sense it is a moot point whether we should view the problem of lexemes like **sail** (*n*) and **sail** (*v*) as a zero derivation or as part of the lexicon whose form class membership is syntactically defined, if both views result in them being handled the same way in the dictionary—as subentries of a single entry.

However, if there is a distinction in the lexicon between, for example, the following categories, then we must indicate each portion of the lexicon as a different category:

Category A —that part of the lexicon that is inherently nominal and must take verbal derivations to function verbally.

Category B —that part of the lexicon that is inherently verbal and must take nominal derivations to function nominally.

Category C —that part of the lexicon that can function in either capacity with either no derivation or with either derivation.[5]

---

[4]Zgusta (1971:80-89) recognizes a vague intermediate status which he calls "partial homonymy" and acknowledges some of the complexities of the issue.

[5]We are sure there are a variety of solutions for different types of languages and regions of the world. One possibility, as suggested in chapter 9, is to use a broader term, such as *relater*, where the membership is more flexible than strictly *preposition* or *conjunction*. Another possibility is to distinguish something like **Headword** *n* (= inherently nominal), from **Headword** *As n* (= flexible membership syntactically defined).

The critical evidence for deciding between different senses of the same word (polysemy) and different words (homonymy) is a corpus of *natural text examples*. Serious lexicography assumes the presence of a large body of natural text, and an ability to cull through those texts to see the range of meaning encompassed by a lexeme and if and how they contrast.[6] Mental searching by itself is inadequate.

How does one decide, for example, that **just**$_1$ 'only' and **just**$_2$ 'fair, morally right' are separate lexemes, whereas **just**$_1$ has several related senses 1) only (*just sugar*), 2) simply, merely (*they'll just have to go home*), 3) exactly (as in British English *she sat just there*)? In working through the following principles, it is wise to get a variety of native speaker judgments rather than simply (just$_1$, sense 2) relying on the intuitions of the compiler.

The process is dynamic. The lexicographer should plan to revise and refine entries that are suspected to involve homonymy and polysemy.

Principles

1) Is there a thread of shared meaning that is acknowledged as shared by native speakers?

2) If the difference is mainly one of different part of speech (**\ps**) and the language has a large segment of vocabulary where part of speech is a function of the syntax (slot in a sentence) rather than of the lexicon (something inherent in the word itself), then consider handling them as different senses of the same lexeme.[7] Consider:

   shower *v.* washing the body standing under running water;
   — *n.* 1) the place used to shower (v). 2) the fixtures used to shower (v).

   jalan [Indonesian] *vi.* go, walk, move;
   — *n.* path, trail.

3) Where a shared semantic thread is not demonstrable, tentatively handle them as separate lexemes (homonyms). It is natural for lexicographers and their team of

---

[6]The computer program FIESTA provides fast, interactive concordance capabilities for a text corpus of the size normally processed by the average linguist or anthropologist. It can be ordered from International Computer Services, Box 248, Waxhaw, North Carolina 28173 USA. This is the same address used for ordering SHOEBOX.

[7]Some commercial English dictionaries have made an editorial policy where different parts of speech are always handled as separate lexemes. But this grows out of a view of language that is often inaccurate when put up against the data, assuming part of speech is something that is inherent in the lexicon. It is also a natural consequence of lexicographers artificially removing words from communicative contexts and isolating them as atomic units to organize in a alphabetical listing. Any four-year-old can see a relationship between, for example, *cook* (v) and *cook* (n).

assistants to hypothesize about shared meanings, but one should have a healthy disrespect or skepticism about accepting folk etymologies.

```
\lx fuka
\hm 1
\ps vt
\ge open
```

fuka$_1$ *vt.* open.

```
\lx fuka
\hm 2
\ps n
\sn 1
\ge mountain
\sn 2
\ge island
```

fuka$_2$ *n.* 1) mountain. 2) island.

4) Assuming shared meaning, different senses tend to have *different lexical networks* as mapped out in the lexical functions (**\lf**). Most lexicographers tend to limit themselves to examining near synonyms with a paraphrase test, which is good, but it need not be limited to synonyms. [**CAUTION**: Having different lexical networks is also true for homonyms, so one must first establish the related meaning.] For example, with **fuka**$_1$ above:

```
\lx fuka
\hm 1
\ps vt
\sn 1
\ge open
\de open, reveal, undo,
    unfasten
\lf Syn = holik
\le open, undo
\sn 2
\ge explain
\lf Gen = prepa
\le speak, say
```

fuka$_1$ *vt.* 1) open, reveal, undo, unfasten. *Syn:* **holik** 'open, undo'. 2) explain *Gen:* **prepa** 'speak, say'.

In the example above, both senses share the idea of *revealing* in 1) things; in 2) knowledge. But if **holik** were substituted, one would not normally interpret it as 'explain', and if **prepa** were substituted one could not interpret it as 'open, unfasten'.

```
\lx epmata
\ps vt
\sn 1
\ge kill
\lf Nug = geba
\le people
\lf Spec = fage
\le spear s.t.
\lf Spec = rasi
\le poison s.o.
\sn 2
\ge extinguish
\lf Nug = bana
\le fire
\lf Spec = skahik bana
\le pull apart logs to
    let fire die
```

epmata *vt.* 1) kill. *Nug:* **geba** 'people'; *Spec:* **fage** 'spear s.t.'; *Spec:* **rasi** 'poison s.o.'. 2) extinguish. *Nug:* **bana** 'fire'; *Spec:* **skahik bana** 'pull apart logs to let fire die'.

```
\lx caan
\ps v
\sn 1
\ge sense
\de hear, listen, sense
\nt Passive 'hear, sense'
    or active 'listen'
\lf SynD = prenge
\le hear, listen [Lisela]
\sn 2
\ge obey
\lf Syn = hai
\le follow, obey
```

caan *v.* 1) hear, listen, sense. *SynD*: **prenge** 'hear, listen [Lisela]'. 2) obey. *Syn*: **hai** 'follow, obey'.

5) Assuming shared meaning, different senses may have *different grammatical or collocational frames*. [**CAUTION**: This also tends to be true for homonyms].

For example, *big* in the sense of 'large' may collocate with either animate or inanimate nouns, whereas *big* in the sense of 'important' tends to be restricted to humans and events.

In Buru, **fuka**$_2$ (above) is interpreted in the sense of 'mountain' when it collocates with the prepositions 'up' or 'upstream', but as 'island' when it is preceded by 'downstream' (and normally followed by the name of the island).

Also in Buru, **emhuka** by itself is interpreted as 'maiden, young (unmarried) woman', but when followed by a clan name it is simply a classifier indicating that the person is 'female human' and asserts nothing about age.

6) There is more likely to be *ambiguity* between different senses of the same word than between different lexemes.

   For example *big rodeo* is ambiguous between the sense of 'large' and the sense of 'important'.

7) Different senses of the same word can represent a metonymic part-whole or generic-specific relationship. The same is not true for homonyms.

```
\lx man
\ps n
\sn 1
\ge adult male
    human [specific]
\sn 2
\ge human [generic]
```

man *n.* 1) adult male human. 2) human.

```
\lx beton
\ps Time
\sn 1
\ge night [part]
\de nighttime, period of
    darkness in the normal
    daily cycle of dark and
    light
\sn 2
\ge day [whole]
\de entire 24-hour cycle. A
    period of time telling
    number of day's travel,
    number of days since
    s.t. happened, etc
```

beton *Time.* 1) nighttime, period of darkness in the normal daily cycle of dark and light. 2) entire 24-hour cycle. A period of time telling number of day's travel, number of days since s.t. happened, etc.

```
\lx bia
\ps n
\sn 1
\ge palm  [generic]
\sn 2
\ge sago  [specific]
\sn 3
\ge paste [part]
```

bia *n.* 1) palm. 2) sago. 3) paste.

Cautions

1) Lexemes can have meanings that are historically related, but which are currently considered different 'words' by native speakers.

   For example, Spanish *caballero* in its technical parse, and historically, meant 'horseman'. Because historically only aristocracy were allowed to ride horses,

caballero developed the additional sense of ‘gentleman’. This term is currently used in limited contexts in many Spanish speaking areas, and to some Spanish speakers it simply means ‘men’s (toilet)’, and the speakers do not think of ‘horseman’, or even ‘gentleman’ when they see the word.

In English, *wrought* is an archaic form that was once productive as the past participle of *work* (as in ‘What hath God wrought [done]?’). Many English speakers today do not think of the term as meaning ‘worked, done’, but rather ‘ornamented’ and almost exclusively limited to *wrought iron*.

2) Perhaps the most common source of homonyms is the assimilation of borrowed words (**\bw**) into the language. For the compiler who is aware of the linguistic history of a region, these may be easy to spot.

```
\lx basa
\hm 1
\ps vn
\ge spicy
```

basa$_1$ *vn.* spicy.

```
\lx basa
\hm 2
\ps n
\ge language
\bw Sanskrit via Malay
    fi:bahasa
```

basa$_2$ *n.* language. *From:* Sanskrit via Malay *bahasa*.

```
\lx beta
\hm 1
\ps vt
\ge connect
```

beta$_1$ *vt.* connect.

```
\lx beta
\hm 2
\ps PRO
\ge 1s
\re I ; me
\de I, me
\bw Malay
```

beta$_2$ *PRO.* I, me. *From:* Malay.

3) Where the vernacular language is genetically related to the national language the differences between loans and inherited vocabulary may be more difficult to unravel. For example, the vernacular language (Buru), the national language (Indonesian) and the regional lingua franca (Ambonese Malay) all belong to the Austronesian language family. Both Indonesian and Ambonese Malay are derived historically from different strains of Malay (B.D. Grimes 1991). Both are sources for loans in Buru. Sometimes the forms can be identified by principles of historical and comparative linguistics, but there should be cautions, in that semantic shifts

can also take place. Both words in each of the following pairs of words have the same ultimate historical source, but one member of each pair has been directly inherited from the parent language, whereas the other member has taken an indirect route.

```
\lx fofo
\ps n
\ge fish_trap
\et *bubu
\eg fish trap
```

fofo *n.* fish trap. *Etym:* *bubu 'fish trap'.

[inherited vocabulary]

```
\lx bubun
\ps n
\ge fish_trap
\bw Malay
```

bubun *n.* fish trap. *From:* Malay.

[borrowed word]

```
\lx fina
\ps n
\ge female
\et *binay
\eg female
```

fina *n.* female. *Etym:* *binay 'female'.

[inherited vocabulary]

```
\lx bini
\ps n
\ge wife
\bw Malay
```

bini *n.* wife. *From:* Malay.

[borrowed word; historical semantic shift]

Pawley (1993:27/4/93) provides an additional caution:

> Polysemy is certainly common but I think dictionaries tend to exaggerate its frequency and, even when it is clearly present, to handle it badly. There is a common ailment of dictionaries that I will dub 'false polysemy'. The worst offenders are bilingual dictionaries. Conventional bilingual dictionaries start with a methodological handicap. Their first obligation is to give *translation equivalents* not definitions. Therefore, for any term in the source language they tend to distinguish as different senses those aspects of the meaning or reference that require a different translation equivalent in the target language. Suppose that the source language A has a term **tal** meaning 'leg (of animal or furniture)', while target language B has no equivalent. Instead B has three distinct terms, meaning 'shank' (leg from knee to ankle, in case of humans), 'thigh or upper leg' and 'supporting rods of chair, table, etc.'. So the dictionary-maker compiles an entry: [emphasis in original]

```
\lx tal
\ps n
\sn 1
\ge shank
\sn 2
\ge thigh ; upper_leg
\sn 3
\ge rod
\de supporting rods (of
    chair, etc.)
```

tal *n.* 1) shank. 2) thigh, upper leg. 3) supporting rods (of chair, etc.).

[false polysemy]

```
\lx tal
\ps n
\ge shank
\de shank, thigh, upper
    leg, supporting timber
```

tal *n.* shank, thigh, upper leg, supporting timber.

[no polysemy]

Pawley (1993:27/4/93) continues:

> While useful for translation purposes, clearly this procedure is liable to give a very distorted impression of the semantics of the source language. My preference is to first seek to provide a unifying definition for the range of meaning exhibited by a word and only admitting polysemy as a last resort. It is perhaps useful to contrast the *inherent meaning* of a form with its *contextual meaning*. Kicking is usually done with one leg but that is a contextual association, not an inherent restriction on the meaning. [emphasis in original]

For additional reading on homonymy and polysemy, refer to Bartholomew and Schoenhals (1983: Ch.10), Landau (1984: Ch.4), Newell (1986:45ff.), Wierzbicka (1980, 1985, 1986, 1988, 1991, 1992, 1992–ms), or Zgusta (1971).

6.4 Semantic categories (\sd, \th, \is)

Tagging semantic categories is useful for a variety of analytical and publication purposes. The discussion here uses semantic domains (**\sd**), but many of the principles are applicable to use of the thesaurus (**\th**) and index of semantics (**\is**) field codes as well. (See §2.1 for preliminary discussion).

Entries containing the desired semantic domain can be easily extracted from the master lexicon through the use of FILTERS in SHOEBOX for studying groups of related words, such as kin terms, body parts, fish names, plant names, carrying verbs, speech-act verbs, etc. For some parts of a language, indicating a semantic class (in **\sd**) may also provide more grammatical information than simply indicating the part of speech (**\ps**). For example, in the Buru lexicon certain generalizations become available by knowing a verb is a cutting verb:

```
\lx hete
\ps vt
\sd Vcut
\ge cut
\de cut into sections for use
\lf Gen = lata
\le cut
\pd -k
```

hete *vt.* cut into sections for use. *Gen*: **lata** 'cut'. *Prdm:* -k.

This information tells us (following C. Grimes 1991) that this entry shares a basic structure with other cutting verbs:

Subject:Actor:agent — DO:*cut* — (Object:Undergoer:patient)
— (uses preposition **tu** + instrument)

What distinguishes one cutting verb from another tends to be differences in manner, typical instrument, typical object, and occasionally typical agent or purpose. A carrying verb looks something like the following:

```
\lx leba
\ps vt
\sd Vcarry
\ge carry_w/pole
\de carry on the shoulder with
    a pole. Includes object at
    one end, objects at both
    ends, or object in the
    middle carried by two
    people
\lf Gen = ego
\le get, take, transfer control
\pd -h
```

leba *vt.* carry on the shoulder with a pole. Includes object at one end, objects at both ends, or object in the middle carried by two people. *Gen*: **ego** 'get, take, transfer control' *Prdm:* -h.

It shares with other carrying verbs the following general structure:

Subject:Actor:agent — DO:*carry* — (Object:Undergoer:figure[8]) —
(preposition **tu** + instrument)
— (preposition **fi di** + locative source) — (preposition **gam di** + locative goal)

Similarly, identifying the semantic class of certain types of nouns tells us (again from the grammar description) how this lexeme should behave in certain constructions (such as in the following example, which in the vocative takes the **–n** suffix, **aman**).

[8]'Figure' is the object whose location is in question. Foley and van Valin (1984) use the term 'theme'. With carrying verbs only one oblique argument is normally expressed—the one most salient to the discourse.

\lx ama
\ps n \sd Nkin \ge F \re father ; uncle \de father

ama *n.* father.

Combinations of semantic domains are possible. For example:

\lx flehet
\ps n \sd Ncult ; Ninstr \ge sago_pounder

flehet *n.* sago pounder.

A suggested starter list of semantic domains is found in Appendix C.

6.5 Handling dialect information

MDF provides several strategies for cataloging dialectal information. But before explaining these strategies it is important to address some broader issues. Firstly, language variation limits communication.[9] Variation with definable clusters of patterns *within* a language normally represent what we call 'dialects'. Different dialects normally have unique patterns of history, language contact, and language use.

A bilingual dictionary normally encodes one primary dialect which is explicitly identified, and may include some subsidiary information indicating how related dialects encode similar semantic concepts as:

1) ***Related forms***: structural variants of the primary dialect.

2) ***Unrelated forms***: different lexical items altogether.

3) ***Forms with different functions/meaning***: Semantic shifts represented by the same lexical item used with slightly different meaning in different dialects.

4) ***Froms with different distributional networks***: similar lexemes used with different collocational, contextual, syntactic, or morphological constraints in different dialects.

For example, American English *advertisement* [advr'taizmnt] carries different stress and vowel quality in Australian and British English [ad'vrtɪzmnt] (#1 above). American English *forest* includes areas filled with unplanted trees, whereas Australian English *forest* implies that the trees were planted (#3 above). American English *supper* implies the meal at the end of the day, whereas Australian English *supper* implies a late evening

[9]This phrasing is adapted from the title of Simons (1979).

dessert, rather than the meal (#3 above). American English *flashlight* has a dialectal equivalent in British and Australian English *torch* (#2 above). But American English also has a word *torch* which implies using flame for light (this suggests #3 above). However, British and Australian English also use the word *torch* with the sense which implies using flame for light as does American English. Thus, *torch* in the two dialects can be said to have the same meaning, different meanings and different distributional networks (#4 above).

To mix all dialect variations into a single amorphous cauldron without identifying a primary dialect and without identifying which dialect the variants belong to is confusing to language learners, misleading to comparative linguists, and disappointing to local users who often want the dictionary to give them a strong sense of "this is us; this is our language!" The mixed dialect approach belongs to nobody and represents nobody. [**CAUTION**: Dialectal variants other than the dialect that is targeted as primary *must be explicitly identified*.]

A complication arises in the multipurpose nature of the lexical database—it is not just a dictionary, but it is a receptacle for cataloging other information as well. Some field workers want to use the lexical database as a place to catalog all known variations among dialects. And of course, the lexical database is the appropriate place to do this, even though it may not be appropriate to print all that information in a published dictionary for certain audiences. Some linguists must catalog dialect variants to appropriately use the Computer Assisted Related Language Adaptation [CARLA] programs for adapting texts from one speech variety into a related speech variety.

MDF is structured on the assumption that one dialect is identified as primary in the introduction to the dictionary. Thus, if no other information is given to the contrary, an entry is assumed to represent the primary dialect. All major dialects should be identified in the general introduction to the dictionary, and a dialect map should be included. If an entry represents a different dialect, that dialect should be explicitly identified in the **\ue** (usage) field bundle. Below are two related entries, the first representing the primary dialect (Masarete—and so is unmarked), and the second representing other dialects (Lisela, Rana—marked in the **\ue** field).

```
\lx apu
\ps n
\ge lime
\re lime ; chalk
\de lime slaked from burning
    seashells and used as an
    ingredient in chewing
    betelnut
\et *apuR
\eg lime, chalk
```

apu *n.* lime slaked from burning seashells and used as an ingredient in chewing betelnut. *Etym:* *apuR 'lime, chalk'.

```
\lx ahul
\ps n
\ge lime
\re lime ; chalk
\de lime slaked from burning
    seashells and used as an
    ingredient in chewing
    betelnut
\ue Lisela, Rana
\bw Kayeli
```

ahul *n.* lime slaked from burning seashells and used as an ingredient in chewing betelnut. *Usage:* Lisela, Rana. *From:* Kayeli.

By itself, however, this pattern of using the **\ue** field does not cross-reference semantically related forms. In the lexical functions fields (**\lf**) described in detail in chapter 7, **\lf SynD** is provided for cataloging dialectal synonyms. In using the **\lf** field bundle for this purpose, the contents of the **\le** field identify the dialect, rather than give the gloss. The minor dialect entry should cross-reference the primary dialect form using the **\cf** or **\mn** field bundles. The examples above are modified below to illustrate these uses.

```
\lx apu
\ps n
\ge lime
\re lime ; chalk
\de lime slaked from burning
    seashells and used as an
    ingredient in chewing
    betelnut
\lf SynD = ahul
\le Lisela, Rana dialects
\et *apuR
\eg lime, chalk
```

apu *n.* lime slaked from burning seashells and used as an ingredient in chewing betelnut. *SynD:* **ahul** 'Lisela, Rana dialects'. *Etym:* *apuR 'lime, chalk'.

```
\lx ahul
\ps n
\ge lime
\re lime ; chalk
\de lime slaked from burning
    seashells and used as an
    ingredient in chewing
    betelnut
\ue Lisela, Rana
\bw Kayeli
\mn apu
```

ahul *n.* lime slaked from burning seashells and used as an ingredient in chewing betelnut. *Usage:* Lisela, Rana. *From:* Kayeli. *See main entry:* **apu**.

Some MDF users are annoyed by this strategy that prints the dialect name in single quotes following the general strategy MDF uses with the **\le** field. Where dialect differences represent different lexemes altogether, using **\lf SynD =** is certainly appropriate. But MDF also provides the **\va** bundle of fields for handling dialectal variants (i.e. **\va, \ve, \vn, \vr**) where the dialectal variant is given in **\va**, **\ve** gives the English version of the dialect

name and/or any pertinent comment, which MDF will print enclosed in (parentheses), **\vn** the national language version of the dialect name or comments, and **\vr** the regional language version of the dialect name or comments. The **\va** (variants) field is dual purpose. It is intended for identifying structural variants or spelling variants in the primary dialect (e.g. **\lx** examination, **\va** exam; **\lx** cannot, **\va** can't; **\lx** aren't; **\va** aint). It can also indicate the forms of other dialects. The following example is from Indonesian:

```
\lx tidak
\ps NEG
\ge no
\re no ; not
\de no, not; standard negation
    targeting the predicate
\lf Sim = bukan
\le negator of nominal
    arguments
\va tak
\ve formal, written
\va seng
\ve Ambonese Malay
\va sonde, son
\ve Kupang Malay
\va tara
\ve North Moluccan Malay
```

tidak *NEG*. no, not; standard negation targeting the predicate. *Sim:* **bukan** 'negator of nominal arguments'. *Variant:* **tak** (formal, written); **seng** (Ambonese Malay); **sonde, son** (Kupang Malay); **tara** (North Moluccan Malay).

There are additional fields that are appropriate to use for clarifying dialectal information. Complex information on semantic differences, social usage, forms, or distribution can be spelled out at length using the **\ns** (notes on sociolinguistics) field. In addition to the **\ue** (usage) field described above, the often underutilized **\oe** (restrictions) field could be used to explain forms that are restricted to certain dialects.

7. Relating headwords to their lexical networks (lexical functions – \lf)

The notion of *lexical functions*[1] allows systematic exploration of the meaning of a lexeme within its culturally associated relationships, and to associate a lexeme with the words and phrases with which a native speaker associates it, regardless of whether or not one form is a morphological derivation of the other, sharing the same root. One can map the emic networks of meaning of a culture as expressed through the language.

The use of lexical functions was pioneered by Apresyan, Mel'chuk, and others who noticed that regular relationships of meaning operate in a different dimension than do structural patterns. The classic example of this kind of relationship is that semantically *drive* relates to *driver* in the same way that *fly* relates to *pilot*, *write* relates to *writer*, and *treat* relates to *doctor*. They are all typically associated as doers of the actions, but note that not all actor nouns use the English *–er* suffix on the verb of the action.[2] These pairs of words are related semantically, and using lexical functions helps us explore and record the networks of lexical associations controlled by the native speaker.

Using lexical functions not only helps us systematically record meaning relationships, but it is also easy to learn a core set of common functions and expand from there. Many language assistants seem to find the approach intuitive. C. Grimes (1987:25) reported on fieldwork in Buru (a Central Malayo-Polynesian language of the Austronesian family, eastern Indonesia):

> We regularly found that after an hour's session with a language helper we would have enough data to keep us working on it for a whole day. Language helpers frequently were not ready to quit when we were, because they were enjoying themselves so much. In many cases, using this system of exploring the language, the following day the language helper would start off adding information he or she had been mulling over from the previous day's session. In one instance, a man whom I would see for only two or three days out of a month whenever I got down to his village, would point out additional information related to what we had explored the month before![3]

---

[1]While known in most of the literature as 'lexical functions', some also use the term 'lexical relations' to avoid the potential for confusion with LFG [Lexical Functional Grammar] with which it has no relation.

[2]Additional actor nouns are also associated with these verbs, but with more specialized senses. E.g. *chauffeur, (navy) flyer, author, nurse,* etc.

[3]Since that article was written I (Grimes) have had a friend walk for two-and-a-half days through the mountains from his village to mine, to tell me follow-up information about some lexical networks we had been exploring together more than a year before when I had lived in his village. He thought it was interesting information that I should know.

J. Grimes (1992:125) similarly reports about his work among the Huichol (a Uto-Aztecan language of west-central Mexico):

> The intriguing thing about following the paths defined by lexical functions is that the informants themselves, even when totally unsophisticated by academic standards, have an intuitive grasp of what is going on and become more and more interested. It was not uncommon for me to have Huichol friends who stopped by casually to see what was going on come back a day or two later after having thought of another lexical correlate, or having remembered a form the rest of us had on the tip of our tongue but couldn't quite remember. I have never seen that level of involvement when working on syntax.
>
> Delayed reaction was normal. After we thought we had exhausted the lexical neighborhood of one word and gone on to another, values of other lexical functions of the first word would pop into people's heads. They would interrupt, and we would go back and fill in. We made it a regular procedure to stop every so often and ask each other, "What else?" It was impossible to simply work our way down a list; we were traveling around and back and forth within semantic neighborhoods most of the time.

The bundle of field markers used for lexical functions (or a subset of them) is found below. They can be inserted as needed in SHOEBOX through the DATABASE TEMPLATE, manually, or through the use of a MACRO.

\lf	[lexical function]
\le	[English gloss of lexeme in **\lf** field]
\ln	[national language gloss of **\lf** field]
\lr	[regional language gloss of **\lf** field]

\lf bundles can be used recursively within a record as needed. Using a limited number of field markers simplifies the formatting for later printing a dictionary—all lexical functions are handled in the same way for printing. Using the FILTERS in SHOEBOX provides for powerful search and retrieval possibilities.[4] The format for using the **\lf** field bundles is as follows:

[4]For example, a filter set up as [lf]Ant] allows one to look at all antonym relations in the lexicon.

```
\lx huma
\sd Ncult; Nhouse
\ps n
\pn kb
\ge house
\re house ; hut ; building
    ; dwelling
\de any building or houselike
    structure for shelter or
    shade
\gn rumah
\lf Group = fenlale
\le village
\ln kampung
\lf Part = heset
\le wall
\ln dinding
\lf Part = atet
\le roof, thatch
\ln atap
\lf Part = subu
\le door
\ln pintu
\lf Mat = kau okon
\le tree bark
\ln kulit kayu
\lf Mat = srahen
\le split bamboo
\ln bambu
\lf Spec = humkolon
\le garden house, grain bin
\ln rumah kebun
\lf Spec = huma endefut
\le residential house
\ln rumah tinggal
\lf Spec = huma braun
\le meeting house
\ln baileo, balai desa
↓
\dt 9/9/90
```

huma *n.* any building or houselike structure for shelter or shade. *Group*: **fenlale** 'village'; *Part*: **heset** 'wall'; *Part*: **atet** 'roof, thatch'; *Part*: **subu** 'door'; *Mat*: **kau okon** 'tree bark'; *Mat*: **srahen** 'split bamboo'; *Spec*: **humkolon** 'garden house, grain bin'; *Spec*: **huma endefut** 'residential house'; *Spec*: **huma braun** 'meeting house'.

Below is a brief listing with description of lexical functions used in the *Encyclopedic Dictionary of the Buru Language* (ms). Additional lexical functions which have been shown to be relevant for languages like Russian or English, but which we have not yet found to be applicable to Buru may be found listed and described in the works of Igor Mel'chuk and Apresyan (various, cited in bibliography), or in J. Grimes (1990, 1992, and ms). Applying them to a specific dictionary project and interaction with language assistants using lexical functions is described in C. Grimes (1987). In several cases a number of Mel'chuk's functions have been generalized under a single lexical function for

ease of learning and use. The abbreviations in Mel'chuk's or J. Grimes' schema are given in square brackets following the description to link the MDF lexical functions with comparable or closely related lexical functions in their systems. The symbol [~] indicates similar to or encompasses.

****

Syn ***Synonym***: Forms substitutable for the headword in most contexts (exact synonyms are rare). [~ Syn, Syn^, Syn<, Syn>]. Some synonyms are more restricted in their collocations than the headword [Syn<], and some cover more territory [Syn>].

```
\lx beka
\ps AUX
\ge first
\de first (before doing
    s.t. else)
\lf Syn = peni
\le first (before doing
    s.t. else)
```

beka *AUX.* first (before doing s.t. else). *Syn*: **peni** 'first (before doing s.t. else)'.

SynD ***Dialectal synonym***: Usually equivalent to headword. Dialect named in **\le** field. Alternatively **\va** (variant) and **\ve** can be used (see §6.5).

```
\lx inhadat
\ps n
\ge mosquito
\lf SynD = senget
\le Rana, Lisela
```

inhadat *n.* mosquito. *SynD*: **senget** 'Rana, Lisela'.

SynL ***Loan synonym***: Loans assimilated into everyday speech (common or frequent usage sometimes having adapted to vernacular phonotactics) which are equated with or substitutable for the headword.

```
\lx ka
\ps TAM
\ge HAB
\de habitual aspect
\lf SynL = jaga
\le Ambonese Malay
    "habitual"
\oe fv:ka tends to be used
    in nominal
    constructions, whereas
    fv:jaga tends to be
    used in verbal
    constructions.
```

ka *TAM.* habitual aspect. *SynL*: **jaga** 'Ambonese Malay "habitual"'. *Restrict:* **ka** tends to be used in nominal constructions, whereas **jaga** tends to be used in verbal constructions

SynR ***Register synonym***: Synonym in another speech register (as in speech levels of Javanese, Balinese, or Sundanese).

```
\lx irung [Javanese]
\ps n
\ge nose
\lf SynR = grana
\le H [Krama Inggil]
```

irung *n.* nose. *SynR*: **grana** 'H'.

SynT ***Taboo synonym***: Usually equivalent, but can also have non-taboo range of meaning that is different. Often lexicalized circumlocutions. More localized than **SynD**.

```
\lx minjangan
\ps n
\ge deer
\lf SynT = wadun
\le deer, (back of neck)
```

minjangan *n.* deer. *SynT*: **wadun** 'deer, (back of neck)'.

```
\lx uran
\ps n
\ge shrimp
\lf SynT = sehe
\le shrimp, (reverse)
\et *uDang
\eg shrimp, lobster
```

uran *n.* shrimp. *SynT*: **sehe** 'shrimp, (reverse)'. *Etym:* *uDang 'shrimp, lobster'.

Gen ***Generic (hyperonym)***: A term that is semantically broader than and subsumes headword. Implies a generic-specific relationship, so it should also be cross-referenced as a specific under the entry for the generic. These should follow native speaker intuitions about what term the headword clusters under. The generic term should always be able to substitute for the specific. One can often elicit or check generics by exploring natural kinds or classes with frames such as 'x is a kind of (generic)', 'x is a type of (generic)', 'x belongs to the (generic) class', 'x is a member of the (generic) class'. [= Gener]. (See §8.1 for a discussion of folk taxonomies).

```
\lx feten
\ps n
\ge millet
\de foxtail millet
\lf Gen = agat
\le grain
```

feten *n.* foxtail millet. *Gen*: **agat** 'grain'.

```
\lx sgege
\ps vt
\ge carry
\de carry under-arm
\lf Gen = ego
\le get, take, carry
```

sgege *vt.* carry under-arm. *Gen*: **ego** ‘get, take, carry’.

Spec ***Specific (hyponym)***: A term that is semantically subsumed under the headword. Types of a kind. Check that these follow the emic groupings, rather than reflecting the lexicographer’s ideas about how native taxonomies ‘ought to be’. All of the known specifics should be listed under the entry for the generic term. These generic-specific relationships should be reciprocally cross-referenced. While not technically consistent with the principles of lexical functions, for convenience some compilers use *Spec* to give a phrasal example of nominal headwords rather than giving a fuller sentence example using **\xv**. [~ Spec, Species, Female, Male, Subadult, Child]. (See §8.1 for a discussion of folk taxonomies).

```
\lx lata
\ps vt
\ge cut
\lf Spec = bisi
\le carve
\lf Spec = hete
\le cut into sections for
    use
```

lata *vt.* cut. *Spec*: **bisi** ‘carve’; *Spec*: **hete** ‘cut into sections for use’.

```
\lx enhero
\ps n
\ge spear
\lf Spec = pangneet
\le six-barbed spear
\lf Spec = pangat goit
\le special spear for
    killing humans
```

enhero *n.* spear. *Spec*: **pangneet** ‘six-barbed spear’; *Spec*: **pangat goit** ‘special spear for killing humans’.

Sim ***Similar***: Near synonyms or other terms at the same level of native taxonomy that are subsumed under the same generic term and are relevant for clarifying the headword. These terms are often given in describing the headword, saying “x is like y, but different.” Normally, the more thorough list of the generic-specific taxonomy should be found under the generic term, rather than listing many *Sim* under each specific. For Buru, reproducing all 17 cutting verbs under each specific entry is not economical. [~ Syn^, Syn<, Syn>].

```
\lx pangneet
\ps n
\ge spear
\de six-barbed spear
\lf Sim = pangpaat
\le four-barbed spear
```

pangneet *n.* six-barbed spear. *Sim*: **pangpaat** ‘four-barbed spear’.

```
\lx bisi
\ps vt
\ge carve
\lf Sim = dasa
\le cut to a sharp point
```

bisi *vt.* carve. *Sim*: **dasa** ‘cut to a sharp point’.

Nact ***Actor noun***: Doer of verb, implying habitual or characteristic association. [~ S1 (first substantive), N1 (first nominal argument)].[5]

```
\lx ekfilik
\ps vt
\ge sell
\lf Nact = gebkaleli
\le merchant
```

ekfilik *vt.* sell. *Nact*: **gebkaleli** ‘merchant’.

Nug ***Undergoer noun***: Typical undergoer of a verb; the undergoer implied if none specified. [~ S1, S2, N1, N2].

```
\lx hete
\ps vt
\ge cut
\de cut into sections for
    use
\lf Nug = kau bana
\le firewood
```

hete *vt.* cut into sections for use. *Nug*: **kau bana** ‘firewood’.

Nloc ***Noun of location***: Location normally associated with headword. [= Nloc].

```
\lx agat
\ps n
\ge grain
\de grain (dried)
\lf Nloc = humkolon
\le grain storage house
```

agat *n.* grain (dried). *Nloc*: **humkolon** ‘grain storage house’.

[5]Using Nact, Nug, Ninst, etc. is a different strategy from the N0, N1, N2, N3 used by Mel’chuk and company. We find our current system far more practical both for remembering and for training others to use *lexical functions*.

Ninst ***Instrument noun***: Instrument associated with the action of the headword; the instrument implied if unspecified. [~ S3, N3].

```
\lx bisi
\ps vt
\ge carve
\lf Ninst = katuen
\le machete
```

bisi *vt.* carve. *Ninst*: **katuen** 'machete'.

```
\lx dihi
\ps vt
\ge comb
\lf Ninst = dihit
\le comb (n)
```

dihi *vt.* comb. *Ninst*: **dihit** 'comb (n)'.

Nben ***Benefactee***: The one who benefits from the activity. The one implied if none specified.

```
\lx soso
\ps vt
\ge nurse
\lf Nben = anmihan
\le infant
```

soso *vt.* nurse. *Nben*: **anmihan** 'infant'.

Ngoal ***Noun of goal***: Typical or unspoken goal associated or implied by headword.

```
\lx oli
\ps vi
\ge return
\lf Ngoal = huma
\le house, home
```

oli *vi.* return. *Ngoal*: **huma** 'house, home'.

Ndev ***Deverbal noun***: [~ S0, N0].

```
\lx iko
\ps vi
\ge go
\lf Ndev = enyikut
\le (his/her) going
```

iko *vi.* go. *Ndev*: **enyikut** '(his/her) going'.

Res ***Result***: Consequence, resulting state or event. [= Res, Conseq].

```
\lx mata
\ps vn
\ge die
\lf Res = enmata
\le death
```

mata *vn.* die. *Res*: **enmata** 'death'.

Whole ***Noun of the whole***: The whole, of which the headword is a part.

```
\lx bubu enitu
\ps n
\ge ridgepole
\lf Whole = huma
\le house, building
```

bubu enitu *n.* ridgepole. *Whole*: **huma** ‘house, building’.

```
\lx maen
\ps n
\ge handle ; shaft
\lf Whole = enhero
\le spear
```

maen *n.* handle, shaft. *Whole*: **enhero** ‘spear’.

Part ***Part of the whole***: The part, of which the headword is the whole.

```
\lx huma
\ps n
\ge house
\lf Part = kasa
\le rafter
\lf Part = subu
\le door
```

huma *n.* house. *Part*: **kasa** ‘rafter’; *Part*: **subu** ‘door’.

Mat ***Material***: Material used to make headword, or material of which it is composed.

```
\lx atet
\ps n
\ge thatch
\lf Mat = bia omon
\le sago palm leaves
```

atet *n.* thatch. *Mat*: **bia omon** ‘sago palm leaves’.

Vwhole ***Verb of the whole***: [~ V0]. This is the converse of **Whole**.

```
\lx enyikut
\ps n
\ge going
\lf Vwhole = iko
\le go
```

enyikut *n.* going. *Vwhole*: **iko** ‘go’.

Serial ***Conventionalized serial constructions*** using headword.

```
\lx heka
\ps vi
\ge move
\de move away quickly
\lf Serial = heka tuha
\le run off with s.o. or
    s.t.
```

heka *vi.* move away quickly. *Serial*: **heka tuha** ‘run off with s.o. or s.t.’.

Compound ***Lexicalized compounds*** using headword.

```
\lx heka
\ps vi
\ge move
\de move away quickly
\lf Compound = hektatak
\le abandon s.t.
```

heka *vi.* move away quickly. *Compound*: **hektatak** 'abandon s.t.'.

Sit ***Situation***: Situations involving headword, or activities typically associated with headword.

```
\lx epkiki
\ps vi
\ge dance
\lf Sit = pesta kaweng
\le wedding celebration
```

epkiki *vi.* dance. *Sit*: **pesta kaweng** 'wedding celebration'.

Prep ***Preparatory activity***:

```
\lx atet
\ps n
\ge thatch
\lf Prep = sau atet
\le sew thatch
```

atet *n.* thatch. *Prep*: **sau atet** 'sew thatch'.

Phase ***Phases of head***: For example, processes of building, making, growing, time cycles, etc. [~ Phase, Seq, Child, Adult].

```
\lx fultimo
\ps Time
\ge east_monsoon
\lf Phase = Samsama
\le lunar month around
    August
```

fultimo *Time.* east monsoon. *Phase*: **Samsama** 'lunar month around August'.

Max ***Superlative degree*:** Intense or extreme degree of headword; the outside limit. 'As x as you can get'. [~ Super, Magn, Incr ('more than last time checked'), Plus ('more than expected')].

```
\lx bana
\ps n
\ge fire
\lf Max = pothaki
\le forest fire
```

bana *n.* fire. *Max*: **pothaki** 'forest fire'.

```
\lx reden
\ps n
\ge dark
\lf Max = reden tuni walet
    mite
\le pitch black
```

reden *n.* dark. *Max*: **reden tuni walet mite** ‘pitch black’.

Min ***Reduced/diminished degree***: Minimized or decreased state of headword. [~ Decr (‘less than last time checked’)]

```
\lx bage
\ps vn
\ge sleep
\lf Min = bagleak
\le nap, siesta
```

bage *vn.* sleep. *Min*: **bagleak** ‘nap, siesta’.

Degrad ***Degradatory degree***: Deteriorated or decayed state.

```
\lx tonal
\ps n
\ge cuscus
\de cuscus marsupial
\sc Phalanger spp
\lf Degrad = mefu
\le rotten
```

tonal *n.* cuscus marsupial. *Phalanger spp*. *Degrad*: **mefu** ‘rotten’.

```
\lx kau
\ps n
\ge wood
\lf Degrad = bono
\le decayed
```

kau *n.* wood. *Degrad*: **bono** ‘decayed’.

Caus ***Causal***: [~ Caus, Perm].

```
\lx emgea
\ps vn
\ge embarrassed
\lf Caus = pemgea
\le embarass s.o.
```

emgea *vn.* embarrassed. *Caus*: **pemgea** ‘embarass s.o.’.

Start ***Inceptive***: Initial phase, inceptive, inchoative. [~ Incep, Prox].

```
\lx bana
\ps n
\ge fire
\lf Start = enhewek bana
\le light a fire
```

bana *n.* fire. *Start*: **enhewek bana** ‘light a fire’.

Stop ***Cessative***: Final phase. [~ Fin (the situation ends), Cess, Liqu (s.o. causes the situation to end), State].

```
\lx dekat
\ps n
\ge rain
\lf Stop = dekat dere
\le rain lets up
```

dekat *n.* rain. *Stop*: **dekat dere** 'rain lets up'.

```
\lx enein
\ps v
\ge work
\lf Stop = deak
\le stop, rest from
    activity
```

enein *v.* work. *Stop*: **deak** 'stop, rest from activity'.

Feel ***Sensation of headword***: In many cases it is appropriate to indicate both the sensation or feeling or symptom of illness and the body part where it is manifested (e.g. tickly nose) [~ Manif (feeling, body part), Sympt (illness, body part)].

```
\lx bana
\ps n
\ge fire
\lf Feel = poto
\le hot
```

bana *n.* fire. *Feel*: **poto** 'hot'.

Sound ***Sound*** uttered by or characteristically associated with headword. [~ Son].

```
\lx dole
\ps n
\ge frog
\lf Sound = troo-troo
\le ribet
```

dole *n.* frog. *Sound*: **troo-troo** 'ribet'.

Cpart ***Counterpart***, complement, or converse (but not antonym). No cultural middle ground or gradation along a process or scale. Concepts like 'more' and 'less' do not apply. For Buru includes male/female, inside/outside names. [~ Conv (permutes arguments formally staging the same transaction from different viewpoints such as with 'buy' and 'sell'), Comp].

```
\lx kete
\ps n
\ge parent_in_law
\lf Cpart = emsawan
\le son-in-law,
    daughter-in-law
```

kete *n.* parent in law. *Cpart*: **emsawan** 'son-in-law, daughter-in-law'.

Ant ***Antonym***: Opposite extreme of a process or scale. 'More' and 'less' apply. [~ Anti, Rev].

```
\lx emhama
\ps vn
\ge light(weight)
\lf Ant = beha
\le heavy (thing)
\lf Ant = emteno
\le heavy (person)
```

emhama *vn.* light(weight). *Ant*: **beha** 'heavy (thing)'; *Ant*: **emteno** 'heavy (person)'

Head ***Head of group***: [~ Cap, Lead].

```
\lx noro
\ps n
\ge kin_group
\lf Head = gebhaa
\le local kin group head
```

noro *n.* kin group. *Head*: **gebhaa** 'local kin group head'.

Group ***Group:*** collective or concentration of headword: [~ Group, Equip, Mult, Organization].

```
\lx fafu
\ps n
\ge pig
\lf Group = fafu reren
\le pig herd
```

fafu *n.* pig. *Group*: **fafu reren** 'pig herd'.

```
\lx geba
\ps n
\ge person
\lf Group = geba rano
\le crowd of people
```

geba *n.* person. *Group*: **geba rano** 'crowd of people'.

```
\lx uka
\ps n
\ge bamboo
\de bamboo (generic)
\lf Group = uka lale
\le stand of bamboo
```

uka *n.* bamboo (generic). *Group*: **uka lale** 'stand of bamboo'.

Unit ***Single unit of headword***: Single piece or occurrence. [~ Sing, Indiv].

```
\lx uka
\ps n
\ge bamboo
\lf Unit = uka walan
\le bamboo pole
\lf UnitPart = uka kasen
\le section of bamboo
```

uka *n.* bamboo. *Unit*: **uka walan** 'bamboo pole'; *UnitPart*: **uka kasen** 'section of bamboo'.

ParS ***Parallelism (same)***: Parallelism attested in formulaic, ritual or poetic text, meaning (in that context) effectively the same as the headword. These associations may not occur in normal speech.

```
\lx saka
\ps DEIC
\ge up
\lf ParS = lepak
\le go up, ascend
```

saka *DEIC*. up. *ParS*: **lepak** 'go up, ascend'.

ParD ***Parallelism (different)***: Parallelism attested in formulaic, ritual or poetic text implying a counterpart, opposite or complementary category to the headword. Like Cpart and Ant, but in formulaic language, often with a sense not found in ordinary language.

```
\lx saka
\ps DEIC
\ge up
\lf ParD = pao
\le down
```

saka *DEIC*. up. *ParD*: **pao** 'down'.

```
\lx supan
\ps Time
\ge morning
\lf ParD = emhawen
\le evening
```

supan *Time*. morning. *ParD*: **emhawen** 'evening'.

Idiom ***Conventionalized expressions*** using headword.

```
\lx agat
\ps n
\ge grain
\lf Idiom = aga lahin
\le inheritance
```

agat *n.* grain. *Idiom*: **aga lahin** 'inheritance'.

For an alphabetized starter list of the lexical functions described in this chapter, see Appendix D.

Users can use this **\lf** bundle to adapt needs not explicitly mentioned in this *Guide.* In other words, users can use the **\lf** bundles to create or *customize* their own categories and labels, keeping in mind that what comes before the equals sign [=] is *italicized* as a label, what comes after the equals sign is assumed to be vernacular and is formatted as such, and what comes in the **\le, \ln,** and **\lr** fields is enclosed in single quotes. For example, one user of an earlier version of MDF working in Africa wanted to use his lexical database to keep track of other words that are phonotactically similar to the headword, easily confused, and mean something else. We suggested using the **\lf** bundles and creating the label 'Not =', with or without the **\le** field as follows:

```
\lx amana
\ps v
\ge gloss
\lf Not = almana, amanna
```

amana *v.* gloss. *Not:* **almana, amanna**.

```
\lx amana
\ps v
\ge gloss
\lf Not = almana
\le gloss of almana
\lf Not = amanna
\le gloss of amanna
```

amana *v.* gloss. *Not:* **almana** ‘gloss of almana’; *Not:* **amanna** ‘gloss of amanna’.

Notice that ‘Not =’ has nothing to do with the concept of lexical functions itself, but it is the formatting sequences of the **\lf** field bundle that is being ‘borrowed’ for other purposes. These **\lf** bundles can be adapted to the needs of the language and the needs of the compiler. Similarly, the **\ee** (encyclopedic) bundle of fields contains no labels or formatting and may be used as a general all-purpose field, not restricted to just encyclopedic information.

8. Considerations for special classes of entries

A common struggle faced by lexicographers dealing with poorly documented languages and cultures is the tension between artificial ideas about a 'pure' dictionary, and the addition of encyclopedic information. The tension between the grammarian's view of the lexicon versus the lexicographer's view of the lexicon is like that of the minimalist versus the maximalist. There are no clear-cut boundaries between the two. On the one hand, factors such as time, lack of authoritative information, editor's demands, and publishing costs weigh against a lot of encyclopedic information. On the other hand, a desire for accuracy, completeness of information, and representing the beauty of the language and culture as interrelated systems, together insist that a certain amount of so-called encyclopedic information be included. The researcher also feels the need to present information that may not otherwise be published,

The **\bb** field (bibliographical reference) is provided to reference ethnographic or other literature which may deal with a subject at greater length. Reference can thus help keep the dictionary succinct but also direct the reader to fuller information elsewhere.

The information in this chapter provides a *starting point* for a number of special types of entries.[1] For flora and fauna, it can take several years to build up a library of useful source books for the region, so the compiler of a lexicon is encouraged to begin early, budget funds, and take every opportunity to purchase good sources. In many universities, both in the dictionary-maker's home country and in the country of their target research, there are capable botanists, ethnobotanists, and zoologists who might be willing to team up with the linguist or anthropologist, accompany them to the field, and complement the lexicographer's local knowledge of the language and culture with their own expertise. In any case, the compiler of a lexicon whose background is in the social sciences should expect to become a self-educated hobbyist in botany and zoology, all the while remembering that they are amateurs.

At least as early as Aristotle (*Categoriae* in McKeon 1941:7–39), it was put forward that a *definition* should be composed of species, genus, and differentiae. In the classical example, *man is a mortal rational animal*, 'man' is the species, 'animal' the genus, and 'mortal rational' the differentiae, or characteristics that distinguish or contrast that species from other members of the same genus.

[1]Many of the ideas in this chapter are adapted with permission from notes and discussions with Prof. Andrew Pawley of the Australian National University, who has been grappling with many of these issues over many years in the course of compiling dictionaries of Kalam, a Papuan language of Papua New Guinea, and Wayan, an Austronesian language of northeastern Fiji.

8.1 Folk taxonomies

When writing the definition or description (**\de, \ee** and **\nt** fields) of a plant or animal, it is helpful to first state what general class or higher category it is a member of—preferably reflecting the generic terms under which it is classed in the vernacular.

```
\lx yoho
\ge civet
\de civet cat; k.o. animal that
    lives on the jungle floor..
```

yoho civet cat; k.o. *animal* that lives on the jungle floor. . .

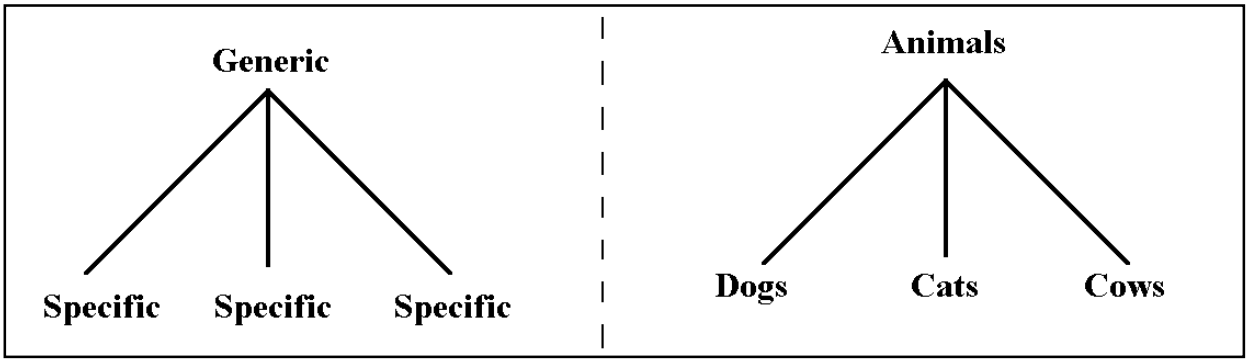
```
\lx bahut
\ge mahogany
\de mahogany; k.o. hardwood
    tree that grows to...
```

bahut mahogany; k.o. hardwood *tree* that grows to. . .

```
\lx pelat
\ge nettle
\de stinging nettle; k.o. shrub
    with leaves spanning...
```

pelat stinging nettle; k.o. *shrub* with leaves spanning. . .

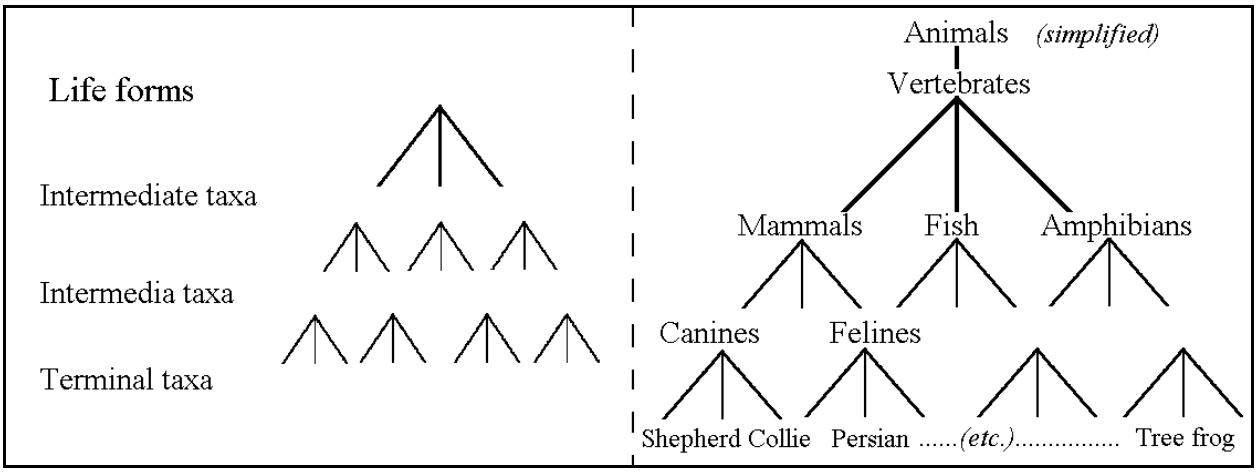
It is a fascinating challenge to become immersed in indigenous systems of terminology or nomenclature, commonly referred to as 'folk taxonomy'. Most languages have at least two levels of a taxonomy (conceptually similar to generic and specific).



But many languages have more complex systems with three or more levels, providing intermediate levels of classification. The nomenclature at the highest (broadest) level of the taxonomy are called *life forms*. The nomenclature at the lowest (most specific) level of the taxonomy are called *terminal taxa* (often popularly referred to as 'species', with a finer level referred to as 'subspecies' or 'varieties'). Between these extremes different languages may have one or more levels of *intermediate taxa*. These intermediate taxa are often tricky to sort out.

Finding the names of the terminal taxa ('what they call x'), while full of hidden pitfalls, is relatively easy compared to finding the intermediate taxa and life forms. Exploring folk taxonomies carefully requires finding the cultural-specific framework within which to ask the appropriate questions. Often questions designed to explore similar things at the same

level of taxonomy are framed as "What are its brothers/cousins/companions?" In many languages it is the noun classifier system which gives clues to the next level of taxonomy.

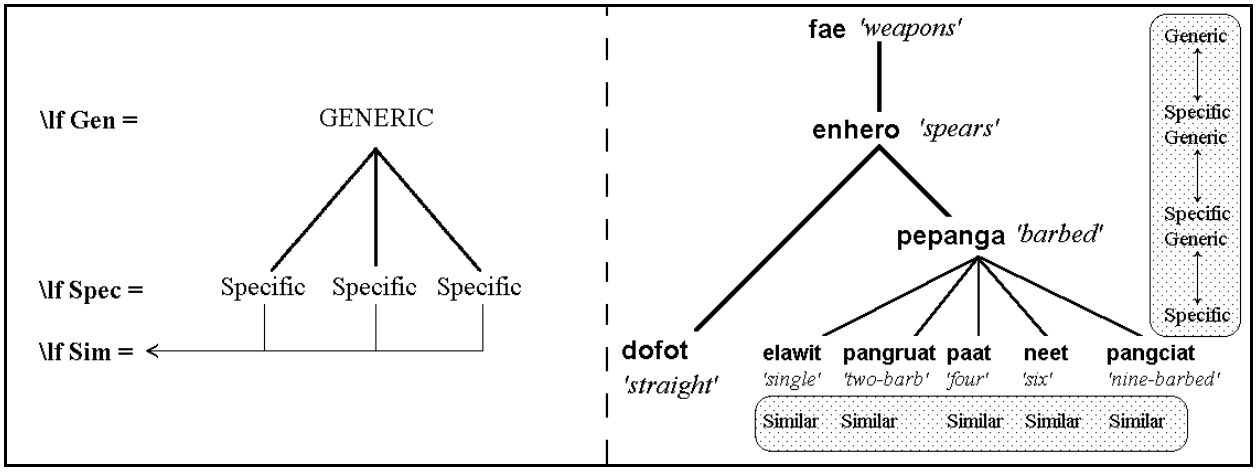


manut flying creatures whose wings are big enough or move slowly enough to see while flying, including birds, bats, and butterflies

man keho *Megapode* sp.
man kumul k.o. large dove...
man tiwit k.o. small bird that feeds on flowers and fruit of *Shorea* trees

man grihit large fruit bat, flying fox
man koi small (10cm body) bat that swoops villages at dusk

In other words, the terminal taxa here involve the classifier indicating the generic term under which these are grouped. In MDF the **\th** field is intended for listing the vernacular generic term under which the **\lx** lexeme is the terminal taxon (see §2.1). Additionally, **\lf Gen =** and **\lf Spec =** are provided for recording the next higher level and next lower level of the folk taxonomy. When a generic term is the headword (**\lx**) all known specifics **\lf Spec =** should be listed in that entry. For entries of each of those specifics, cross-reference back to the appropriate generic with **\lf Gen =**. It is not economical to cross-reference all other terminal taxa that group under the same generic term for each terminal taxa, but **\lf Sim =** is provided to list those that are directly relevant to the headword (see §2.2 and chapter 7).



Cautions:

1) The semantic range of *life forms* between any two languages is rarely isomorphic, particularly between unrelated languages. For example, the kinds of things covered by the Selaru term **masy** is not a direct equivalent of its English gloss 'fish'—the Selaru term includes the English fish, dolphins and whales (which technically are not 'fish', but popularly are) and for some Selaru speakers can include certain mobile shellfish such as lobsters, and perhaps sea slugs.

2) In many languages, *intermediate taxa* and *life forms* may be expressed as verbal propositions (e.g. 'those things that retract their claws', 'those things that have roots'), rather than as simple generic nominals (e.g. 'felines' and 'plants').

3) The system of folk taxonomy may have some, or little correlation with the scientific taxonomy, so one should not expect a good match. This is because the scientific taxonomy is built primarily around similarities and differences in physical structures, whereas the folk taxonomies may put behavioral patterns, or a different physical feature into greater salience for structuring their taxonomies, particularly at the intermediate and life form levels of the taxonomy. That does not make one system better, or the other worse—they are simply different. But to an English or academic audience, the point of reference to identify native flora and fauna through the native nomenclature is the scientific nomenclature. In other words, the lexicographer must identify the native 'emic' system and terminology with reference to the scientific 'etic' system and terminology.

4) Just because what is covered by one native term is handled by two or more scientific terms, or vice versa, does not mean the native community is unaware of the physical similarities and differences in the 'species'. Thus, there may not be great discrepancies in *conceptual correspondence* in many of the terminal taxa (the plants and animals we see—the 'species') between the scientific system and the folk system, but there may be large discrepancies in the *terminological*

correspondence between the two systems. The development of their native taxonomy has simply chosen to make other issues more salient in decisions about 'same' or 'different' at higher levels. For example, the folk taxonomy may have different lexemes for the adolescent (immature) phase, the adult (mature) phase,[2] the male variety, and the female variety, all of which are included under one scientific term (much as we say *foal, colt, yearling, mare, stallion* all referring to *Equus caballus*). Furthermore, the native community may be aware of issues about which the scientific community is unaware. They may explain, for example, that "yes, those two birds are similar in the way that you say, but variety x lives only at the high elevations and feeds on beetles, whereas variety y lives in the jungle lowlands and feeds on grubs in rotten wood." Furthermore, you may be working in a local area where botanists and zoologists have not yet done extensive work, although you should be on the lookout for source books on the region.

5) A number of scholars have observed that flora and fauna of *high cultural significance* tend to be over-differentiated in their terminology. Thus, plants which are intensively cultivated locally (yams, sweet potato, taro, rice, corn, cassava, millet, barley), and animals which are hunted or domesticated and play an intense role in economics, bridewealth, death, clan totems, or religion (pigs, cows, chickens, cuscus, buffalo, water buffalo, etc.), tend to have enriched lexical networks.

6) While the scientific nomenclature and taxonomy is predicated ideally on a principle of distinctive features (much like phonology, in principle), in which one feature is seen as most distinctive or salient in distinguishing one variety from a similar variety, many parts of a native folk taxonomy may define such categories by a convergence of *multiple criteria*. These criteria may include habitat, behavior, potential for eating, and ceremonial significance, as well as physical characteristics such as size, color, texture, pattern, and shape.

7) It is quite common in the world's languages for lexemes to be described with reference to the next highest level of the taxonomy, rather than jumping directly to the highest level. This parallels how part-whole relationships of complex structures work. For example, *toe* is usually described with reference to *foot*, *foot* with reference to *leg*, and *leg* with reference to *body*, rather than *toe* being described directly with reference to *body*. When thinking about the *terminological system* of the folk taxonomy, one must also understand when it is appropriate to refer to life forms and when it is appropriate to refer to intermediate taxa.

---

[2]Use **\lf Phase =** to relate these forms. See §7.

8) No single individual in a society is likely to know all the information sought, so it is useful (i.e. a good technique) to look at and discuss flora and fauna with a core group of native speakers to gain their *collective knowledge*. Using books that have pictures of the flora and fauna in question is helpful, but one must always check whether what they have in mind is identical or slightly different. Such group exploration is fun, and often produces a wealth of new lexemes and new insights that will take additional hours to manage in the lexical database.

Further reading: See Berlin, Breedlove and Raven (1966, 1973, 1974), Bulmer (1967, 1970), Casagrande and Hale (1967), Conklin (1962), Frake (1962), J. Grimes (1980a,b), Lakoff (1987).

8.1.1 Plants

There are some features of use that give a particular plant relevance or prominence in a culture and these should be noted, where found. However, use by itself is not sufficient information for an outside user of the dictionary to identify the particular plant. The dictionary maker must eventually choose which information is most relevant for the published dictionary, but in SHOEBOX using the MDF codes, *all* available information can be recorded and organized for later selection. The following are issues to be considered:

Physical characteristics about the plant's appearance[3]

1) What is the average height of the mature plant?

2) What is the average size (of the trunk, leaves, flowers, fruit)?

3) Is there a distinctive shape or texture (of the trunk, bark, leaves, flower, fruit)?

4) What kind of flowers and fruit does the plant bear, if any? Do these have distinctive color, smell, or taste? [Also list as \lf Part =].

5) Can someone be trained to make an accurate sketch of the plant including detail of the leaves, flower and fruit? [Use **\pc**].

Normal habitat, growth patterns and associated care

1) Does the plant grow wild, is it planted, or both?

2) Where does this variety grow? In the distant gardens, or on the edge of the village? Near the ocean, or inland? In the lowlands, mountains, or coastal plains? In the

[3]Use meters and cm, rather than vague and relative terms such as 'tall', 'large', 'small'.

deep jungle, at the edge of clearings, or in grasslands? Is it associated with a particular kind of soil? [Use **\lf Nloc =**].

3) If it is planted or cultivated, does it need special tending such as stakes for support, weeding, or pruning?

4) When is it planted? When is it harvested? If it is wild, when does it mature or bear fruit?

Uses associated with the plant

1) Is part of the plant used to make something? For example, is the wood used for fence posts, house posts, rafters, bows, spears, firewood, or tools? Is the inner bark used to tie things? If it is a vine, is it used as rope? Are the leaves used as plates, for wrapping, for thatch roofing, or for weaving baskets or mats? Is the bark used to make cloth or string? Are parts of the plant useful for making gourds or buckets?

2) Is the plant (or part of it) eaten? If so, which parts are eaten? Is it eaten raw or cooked? If it is cooked, are there special instruments, materials or preparation needed? Is it cooked with certain other foods? Is it eaten with certain other foods?

3) Does the plant (or part of it) have other uses besides as food and utensils, such as for medicine, oil, poison, glue, dye, perfume? Is it the leaves, the inner bark, the sap, the roots, the fruit or the flowers that are used? How are these prepared?

4) Is the plant used for decoration?

Social values and associated activities

1) Is there a special social value associated with the plant? For example, is it fit for presentation to nobility, or is it eaten only during famine when other foods are not available?

2) Is there special symbolism associated with the plant that requires its presence at certain ceremonies? For example does it symbolize cool things, peace, prosperity, longevity, promises?

3) Does the plant function as a totem that is emblematic of a certain social group?

4) Are there prayers or incantations associated with proper preparation of the plant?

5) If it is planted, do both men and women plant it, only one sex, or are the different sexes involved in different phases of the planting?

6) Do culturally important animals nest in it or under it?

Varieties

1) Are there several kinds of this plant? Do they each have distinct names? Under the most appropriate generic term list the varieties with at least one distinctive feature. [Use **\lf Spec =**]. Use the JUMP feature of SHOEBOX to create separate entries for each of the varieties, also cross-referencing the generic term. [Use **\lf Gen =**].

2) Are there other names for the same plant? [Use **\lf Syn = ; SynD = ; SynR = ; SynT =**].

3) Are there special lexemes associated with phases or stages of this plant's growth? [Use **\lf Phase =**].

8.1.2 Animals

Distinctive physical characteristics of the animal's appearance

1) What is the average size of a mature animal?

2) What is distinctive about the animal's shape or coloring?

3) What are the differences in size, shape, color, or other aspects of appearance between males and females? Between infants, adolescents, and adults?

4) Does it move in a distinctive way?

5) Is there a picture that can be included? [Use **\pc**].

Habitat, growth, and behavioral habits

1) Is the animal wild or domesticated?

2) Is it native or introduced (not native to the area)?

3) Where does it live? In the water or on land? In swamp, jungle, or grassland? In trees, on the ground, or under the ground? On the coast or in the mountains?

4) Does it make a nest, burrow, or find or make shelter in other ways?

5) In what form are the young born (i.e. in eggs or alive)?

6) How many young per birth?

7) Do the parents look after the young? Which parent?

8) Is the animal present year round, is it seasonal, or does it appear only occasionally during times of drought or major storms in other areas?

9) What does it feed on?

10) Does it have a characteristic call, or cry in a distinctive way? [Use \lf Sound = (ribet)].

11) Does it have a characteristic smell?

12) Is it poisonous or aggressive, or otherwise dangerous to people?

Uses

1) Is it eaten by people?

2) Is it fed to other animals?

3) How is it prepared or cooked?

4) Are parts of it used for other purposes? E.g. are its skin, bones, sinews, milk, blood, eggs, horns, fur, or feathers useful?

5) Is the animal used for other purposes? E.g. is it used for hunting, herding, carrying, or pulling heavy loads? Is it kept as a pet?

6) If it is domesticated, how is it raised? If it can be tamed from the wild how is that done?

7) If it is hunted, how is it caught? Note that for culturally important animals there may be many ways. Are special implements used?

Social values and associated activities

1) Are there special beliefs about this animal? E.g. when some animals behave in certain ways they are thought to be omens; some societies believe that certain animals can turn into humans and vice versa; in some societies snakes are associated with evil or with spirits, whereas other societies consider them to represent wisdom, or shrewdness.

2) Are there taboos associated with this animal, or restrictions associated with killing or eating it? Are there avoidance patterns associated with saying its name? Do these types of taboos apply to society at large, to only certain segments of society, to certain individuals, or to certain locales?

3) Does the animal have special value? For example as a totem, in ceremonies, for serving honored guests. For example, in Buru the head of a wild pig or cuscus is given to an honored guest or belongs to the successful hunter. For domestic pigs, a plate full of large cubes of pure pig fat is given to honored guests, whereas plain meat is for the common man.

4) Is the animal considered a pest? Are there special activities for dealing with this?

5) Are there commonly known fables associated with this animal? Does the fable explain prominent physical characteristics or a characteristic call (e.g. 'that is why x has a short tail')?

Varieties

1) Do males and females have different names? [Use **\lf Male =** (stallion, boar, bull); **\lf Female =** (mare, sow, cow)].

2) Do infants, adolescents and adults have different names for different stages of maturity? [Use **\lf Phase =** (lamb, calf, piglet, puppy)].

3) Are there different kinds (varieties) of this animal encompassed by a single term?

4) Are there other names for the same animal? [Use **\lf Syn = ; SynD = ; SynR = ; SynT =**].

8.1.3 Birds

The guidelines for birds are generally the same as for animals above, but particular attention should be paid to:

1) Special patterns or markings on the feathers.

2) Special feeding habits.

3) Restricted ranges of habitat.

4) Special nesting behavior.

5) Special mating behavior.

6) Special calls, particularly those that are characterized by their own lexeme. Some birds have a variety of calls.

7) Special cultural significance. For example, the call of certain jungle birds may be associated with time to get up before dawn; others with the spirits of the dead (as the hoot of an owl in Europe).

8) Myths or fables explaining their call, their appearance, or their behavior.

8.1.4 Fish

The basic guidelines for fish are the same as for animals in general above, but paying particular attention to:

1) Habitat: freshwater, saltwater; river source, deep pools, river mouths; clear water, murky water; tidal pools, surf, reefs, rocks, sandy bottom, deep ocean.

2) Fin structure.

3) Feeding habits.

4) Unique coloring or camouflage.

5) Spawning habits and habitat.

6) Unique ways and instruments used to catch them.

7) Special cultural significance as food, or as totems, or as spiritual intermediaries.

8) Myths or fables explaining their appearance, their behavior, or their significance in other ways.

8.1.5 Insects

The basic guidelines for insects are similar to those for animals in general above, but paying particular attention to:

1) Number and length of wings and legs.

2) Different phases of growth and if the local culture associates, for example, the soft grub growing in the rotten log with the hard-shelled beetle that eventually emerges, or associates the caterpillar with the cocoon, with the butterfly.

3) Which insects are normal food, which are famine food, and which are never eaten. How are they collected, processed and cooked?

4) Are certain insects used as bait for fish or birds?

8.1.6 Body part terms

In making entries for body part terms for bilingual dictionaries, it is particularly easy to mislead the naive reader. In Buru, for example, **kada-n** could be glossed as 'leg', but actually includes both the English 'leg' and 'foot'. Similarly, **faha-n** 'arm' includes both the English 'arm' and 'hand'. Secondary senses or polysemy must be closely scrutinized. For example, Buru **olo-n** 'head' sort of parallels the English, but not quite. To get the sense of 'head of a social group', Buru requires a different morphological structure as **olo** or **pyolot**. Similarly *mouth* of river/jar, *foot* of the hills, *eye* of the storm, *leg* of a journey, deal a *hand*.

1) What is it part of? For English the point of reference is the next larger body part, rather than the whole. For example, *finger* is made with reference to *hand* rather than to *body*. [Use \lf Whole =]

2) Where is it? What does it attach to? Where does it start and end?

3) What other important parts are contained within this part? For example, the *head* includes: eye, ear, nose, mouth, hair, forehead, cheek, chin, temple, brain, etc. The *mouth* includes teeth, tongue, gums, and in some languages lips. [Use \lf Part =].

4) What is it used for? For example, *teeth* are for biting and chewing.

5) Does this body part term apply to both humans and animals? Does it extend to fish and insects?

6) Are there social values or avoidances associated with this body part?

7) Is it valued for food, or for making certain instruments?

8) Are there idioms associated with it? Do these idioms reflect slang or normal speech? [Use \lf Idiom = (he's got a hole in his *head*; he's *foot*-loose and fancy-free)].

9) Is there a picture that can be included (where socially appropriate)? [Use **\pc**].

8.1.7 Kin terms

Kin terms require special consideration for a number of reasons. They are members of a highly structured system. They normally imply certain behavior patterns with links to other members of the system.

1) Is the term (**\lx**) a term of reference (talking *about* s.o.) or a term of address (talking *to* s.o.), or may be used in both ways?

2) If it is a term of reference, is there a lexical or grammatical counterpart for the term of address?

3) Is there a special reciprocal form involving this term and another? For example:

```
\lx feta
\ps n
\sd Nkin
\ge sister_(m.s.)
\lf Group = feta-sar-naha
\le reciprocally brothers and
    sisters, referring to same
    generation males and
    females of different kin
    groups that link to a
    common grandparent
```

feta *n.* sister (m.s.). *Group*: **feta-sar-naha** 'reciprocally brothers and sisters, referring to same generation males and females of different kin groups that link to a common grandparent'.

```
\lx dawe
\ps n
\sd Nkin
\ge WB
\re wife's brother
    ; brother-in-law
\de wife's brother,
    brother-in-law
\lf Group = tal-dawe
\le reciprocally brothers-in-
    law, referring to men of
    different fv:noro who have
    married each other's
    sisters
```

dawe *n.* wife's brother, brother-in-law. *Group:* **tal-dawe** 'reciprocally brothers-in-law, referring to men of different **noro** who have married each other's sisters'.

4) Are there variants or modifications of basic kin terms? For example, are there ways to specify male and female forms?

```
\lx opo
\ps n
\sd Nkin
\ge PP ; CC
\re grandparent ; grandchild
\de grandparent, grandchild;
    signifies plus two or
    minus two generations
\lf Male = opomhana
\le grandfather, grandson
\lf Female = opolfina
\le grandmother, granddaughter
```

opo *n.* grandparent, grandchild; signifies plus two or minus two generations. *Male:* **opomhana** 'grandfather, grandson'; *Female:* **opolfina** 'grandmother, granddaughter'.

5) Can the kin term also be used in a verbal form, or in an extended sense? Consider English *he fathered another child, she mothered too much*. 'Child' in many

languages can also have the sense of 'part of X' or 'diminutive X', often in a genitive construction.

6) Definitions need to accurately encompass the range of meaning and usage of the headword. Translation equivalents are dangerously misleading. Consider **ama** glossed as 'father' (but which actually includes all males of the first ascending generation to ego in the clans of either parent); or **ina** glossed as 'mother' (but which actually encompasses all females of the first ascending generation to ego in the clans of either parent); or **anat** glossed as 'child' (but which actually includes all offspring of the first descending generation to ego's classificatory brothers and sisters).

7) What cultural or behavioral information should be included? Pawley (1993:21/4/93 lecture notes) observes:

> Kinship relations carry a heavy cultural burden. Being a proper mother, father, wife, husband, son, brother, sister, etc. carries certain responsibilities and duties, certain privileges and rights, certain ways of behaving. Should these things be included in the definition or appended to it? I think they should be. It is true that 'mother' in its focal sense, is partly a biological concept. But it is also a social status. Part of the meaning of mother is all the cultural baggage that is associated with this role. Because the social roles differ from one society to another—e.g. in some places brothers and sisters should avoid each other, while in others they can talk and joke freely—these can't be taken as 'givens', they must be spelled out. Obviously, the description must be brief, just an outline of key points, ideally a reference to an ethnography which describes them more thoroughly.

8.1.8 Cultural items (artifacts)

Cultural items (things made from the material world) include such things as houses (of various sorts), gardening implements, weapons, cooking utensils, hunting instruments, cloth, heirlooms, trade items, and objects used to interact with the spirit world or to perform healing and other rituals. Special attention needs to focus on the following:

1) What material are they made from? [Use \lf Mat =].

2) Are they made by everybody, or by specialists?

3) What are they used for?

4) Who uses them, and under what circumstances? Who does not or may not use them, due to cultural norms or cultural taboos?

5) Are there special rituals associated with the objects?

6) Do they have counterparts in the non-human cosmology?

7) Are they involved in ritual or commercial exchange? For example, there may be mats, or cloth, or cooking pots that are exchanged in one direction by certain kin relations at marriage or death. If they are used in ritual exchange, are there other items that are always used to reciprocate in counter-exchange? What are they?

8) Are there metaphors built around, or associated with these items of material culture?

8.1.9 Natural environment

When exploring the natural environment there may be native taxonomies with generics and specifics, such as the more generic *rock* and specifics such as *granite, coral, sandstone, obsidian, limestone, chalk, marble*, etc. There are also different types of clouds, streams, winds, rain, mountains, and constellations.

1) What does it look like? Check for size, shape, color, and texture.

2) Does it have a characteristic smell or taste?

3) What does it feel like?

4) What is it used for?

5) Where is it found?

6) Does it move?

7) What is it like, and what does it contrast with?

8) Are there *kinds* of X? Are there other ways of referring to X?

9) Can it be owned or possessed by a person?

10) Are there animals or creatures or spirits that dwell there?

11) Are there cultural or economic values associated with X?

8.2 Syntactic classes

These are treated in greater detail in chapter 9. Two types are isolated for discussion here: activities and events, and states and processes.

8.2.1 Activities and events

Activities and events characterize actions that are initiated by an Actor, often by a volitional agent. An event encompasses a complex series of activities (such as a wedding, a feast, a litigation, or a trip) which have a definable onset, peak and coda. Different phases of an event may have their own lexemes.

1) Who normally does this activity? Is the action normally associated with a restricted segment of society, such as men, women, young girls? [Use \lf Nact =].

2) What *undergoer* (patient) is assumed if not expressed?

```
\lx hete
\ps vt
\sd Vcut
\ge cut
\de cut s.t. into sections for
    intended use
\lf Nug = kau bana
\le firewood
```

hete *vt.* cut s.t. into sections for intended use. *Nug:* **kau bana** 'firewood'.

3) What *instrument* is assumed if not expressed? [Use \lf Ninst = (machete)].

4) What *location* is normally associated with the activity? [Use \lf Nloc = (jungle, village, gardens)].

5) What *preparatory activity* is necessary before the action can be done? [Use \lf Prep =].

   \lx smoke (meat)
   \lf Prep = cut (meat) into strips

6) What *resulting* thing or state is produced by the action? [Use \lf Nres = ; \lf Result = (cooked)].

8.2.2 States and processes

States and processes characterize qualities, characteristics, state-of-affairs, resulting states or change-of-states that involve a single core argument Undergoer, often by a fully affected patient or an experiencer. See also §9.3.3.

1) Who or what normally is characterized as having this quality or characteristic? Is the characteristic normally associated with humans, non-humans, a restricted semantic domain, or a restricted segment of society, such as men, women, young girls? [Use \lf Nug =].

2) Do these states that represent a 'BE x' relationship have lexical, morphological, or periphrastic causative forms that transform them into a 'BECOME x' relationship? [Use **\lf Cause =**]. A lexical causative is exemplified by the semantic relationship between *big* (BE *big*) and *grow* (cause to BECOME *big*), between *good* and *fix* (cause to BECOME *good*), or between *well* and *heal*. A morphological causative is represented by *wide* and *widen*. A periphrastic causative is represented by *be thirsty* and *make thirsty*. What is expressed in one language as a morphological or periphrastic causative may be expressed in another as a lexical causative.

3) Does this state or process have a special form or idiom for representing an emphasized or an extreme degree? Consider, for example, *black, very black, jet black* (of things), *pitch black* (of the surrounding environment). Notice that the last two represent an 'extreme degree' of black and can be handled with **\lf Max =**. In some languages the *very black* relationship is expressed by a normal adverb, or by reduplication, or an affix. These can be handled in the grammatical introduction if they fit the normal paradigm. If they are unpredictable, the form indicating an intensified degree should be mentioned in the entry.

8.3 Loans and etymologies

Some national language dictionaries do not indicate the source of words even if they are known, for political reasons, publishing economics, or professional insecurity. Or it may be thought that as long as they are assimilated into the language in current usage it is irrelevant whether the lexeme is inherited or borrowed. But for a number of audiences and a number of purposes, if the information is known, and if it is accurate, it is useful to publish. Among the most common users of dictionaries of lesser known languages are comparative linguists who are looking for data. Often, because they do not know or understand the local situation's contact history, they jump to the wrong conclusions and use the wrong data to make a point. Any competent work the lexicographer can do to alert them to what is inherited vocabulary and what is borrowed, not only strengthens the compiler's own perspective but also makes a stronger contribution to how the languages in the region are understood in relation to each other, and through time.

The terms 'loan word' and 'borrowed word' are both misleading and amusing (what language, having once 'borrowed' a word, intends to give it back?). Nevertheless, the terms are fully conventionalized and fully understood by educated audiences as representing vocabulary that has come from another language source, usually due to historical and linguistic contact, and is not directly inherited from the parent language. MDF uses **\bw** for borrowed words. It takes time to become aware of many patterns of borrowings and to correctly identify the source. A common assumption is that borrowed words come from Indo-European colonialist languages. They do, but they also come from neighboring languages, from lingua francas that may have been in the area before the

arrival of any Europeans (such as Swahili, Quechua, Nahuatl, or Ambonese Malay), or long gone languages of no-longer existent empires that once ruled the area. Furthermore, some of these source languages can be genetically related to the language being cataloged.

For historical reconstructions one should be careful to cite attested published reconstructions only in the **\et** field. Use **\nt** or **\ec** field to posit your own guess at a reconstruction. There is a whole science to the principles and procedures of comparative and historical linguistics, and simply trying to work from what looks obvious can quickly get one mired in muck.

Give the gloss of the reconstructed form in the **\eg** field so that the semantic consistency or shift can be seen. Reconstructed meanings for many language families are given in English. Give the original published gloss—do not translate the published reconstructed gloss into the national language, or even into English, as that introduces an extra filter.

The source of the reconstructed form is kept track of in the **\es** field. This is a housekeeping field for data management, not intended for printing. Being able to track down the source of the reconstruction becomes important when analysis begins on these **\et** bundles, because different scholars who do the reconstructions operate on different assumptions, sometimes different principles and methodology, and usually on different data. It eventually becomes apparent which ones are consistently hasty, which are consistently sloppy, and which are consistently meticulous and solid, and therefore reliable or unreliable as the case may be.

Relevant comments can be placed in the **\ec** field, where the connection between the headword and the reconstructed form is not straightforward, where metathesis has occurred, and where there are unexpected sounds, loss, or semantic shift. This field may also be used to posit tentative, unattested reconstructions and supporting data.

8.4 Handling ritual speech and other special registers

Many languages have special speech registers or special uses of lexemes in ritual speech. Javanese, Balinese, and Sundanese, for example, have different 'levels' depending on the social relationship or social posturing between the speaker and addressee.

The linguistic strata of Javanese are perhaps the best described of any special registers in Austronesian languages, particularly the 'high' Javanese *Krama* [kromo] (Poedjosoedarmo 1968, Horne 1974, Wolff and Poedjosoedarmo 1982, Clynes 1989). *Krama* is used to address someone who is socially higher than the speaker "implying a formal or somewhat distant relationship between the speakers" (Horne 1974:xxxi). It is also used in a formal or ceremonial context, such as speech-making at weddings. Javanese speech levels are described by Wolff and Poedjosoedarmo (1982:4):

> Javanese speech levels can be divided roughly into three: the highest, called Kromo; the lowest called Ngoko; and a middle level called Kromo Madyo or just Madyo. There are no clear boundaries between these levels, and Madyo is a continuum between Kromo and Ngoko. The highest level, Kromo, is the refined level, marked by a special vocabulary of somewhat more than a thousand items and a few affixes for which there are special Kromo variants. Kromo is employed to persons of high status. . . Ngoko is the unrefined level with which speakers choose to address persons with whom they are familiar and persons who are not of high status. Ngoko is marked by use of non-Kromo forms for the 1,000 or so items for which there are special Kromo variants.

Horne (1974:xxxi–xxxii) adds:

> The vast majority of Javanese words are neutral with respect to social connotation. But a thousand or so of the most commonly used words in the language are restricted to particular situations defined by the relationship between speakers and the people they are talking about. For each item with built-in social limitations there is at least one other item with the same denotative meaning but complementary social implications. . . . The basic style [register] is Ngoko: there is a Ngoko word for everything, and the Ngoko lexicon is numbered in the tens of thousands. The formal style, Krama is the second largest category having around 850 lexical items. In a Krama-speaking situation, one replaces the neutral (Ngoko) lexicon with Krama vocabulary items when they are available.

There are further substrata within the registers mentioned above. Kromo is additionally marked by precise diction and slightly marked intonation. Nothofer (1982:291) notes that Kromo and Kromo Inggil vocabulary "shows less dialectal variation" than Ngoko.

Buru has a special taboo register that is spoken in the part of the jungle called Garan that has no villages, but takes two days to traverse by foot. In that region the taboo is that nobody is permitted to speak the Buru language, hence the development of this entire special register called *Li Garan* 'the language of Garan'. Most functors are the same as in the common register, but many lexical items (nouns and verbs) have Garan-register forms. These follow the phonotactics of the Buru language, but are different. For example, Li Garan **em-kise-n** 'person, man' replaces the common Buru **geba** 'person, man'. **Kise** normally means 'growing bald' or 'having a high forehead'. The special language of Garan is described more fully in C. Grimes (1991) and Grimes and Maryott (1994).

Speech registers, such as those in Javanese or Buru can be handled just as what was described for handling dialect variation in §6.5. Thus, one can use the variant fields (**\va, \ve, \vn, \vr**), the usage fields (**\uv, \ue, \un, \ur**), the restrictions fields (**\ov, \oe, \on, \or**), the notes on sociolinguistics field (**\ns**), and here, the **\lf SynR =** (register synonym) rather than **\lf SynD =** (for dialectal synonym).

Many languages use *parallelisms* in formal speech, ritual speech, poetry, ballads, or prayers. These parallelisms tend to be of two basic types: the second member of the pair means essentially the *same* as the first member (in this context), as is common in Biblical Hebrew. Or the second member means approximately the *opposite* end of a scale from the first member. These are provided for as **\lf ParS =** (same), and **\lf ParD =** (different). See Fox (1971, 1974, 1975, 1977, 1982, 1988) and Moore (1993) for more discussion and examples. An example of parallelism meaning the same is from a Rotinese poem (Fox 1982:313)

Te leo mafo ai-la hiluk	*But if the trees' shade moves*
Ma sao tua-la keko	*And the lontars' shadow shifts*
Na, Suti, au o se	*Then I, Suti, with whom will I be*
Ma, Bina, au o se	*And I, Bina, with whom will I be*
Fo au kokolak o se	*With whom will I talk*
Ma au dede'ak o se	*And with whom will I speak*
Tao neu nakabanik	*To be my hope*
Ma tao neu namahenak?	*And my reliance?*

An example from Buru that mixes same and different in parallelisms describes hunting a wounded pig.

Kami iko lepak	*We ascended*
iko logok	*we descended*
hama saka	*we searched high*
hama pao.	*we searched low.*

A better known example is based on parallelism from Biblical Hebrew in Psalms 139:7–10 (Jerusalem Bible).

Where shall I go to escape your spirit?
Where shall I flee from your presence?
If I scale the heavens you are there,
if I lie flat in Sheol, there you are.

If I speed away on the wings of the dawn,
if I dwell beyond the ocean,
even there your hand will be guiding me,
your right hand holding me fast.

In many languages which words can pair with which other words is conventionalized in a frozen or semi-frozen state, such that not just any two words can go together. For example, in the Buru example above, **lepak** 'ascend' and **logok** 'descend' pair together as opposites, but **lepak** 'ascend' and **pao** 'down' cannot. These distinctions should be recorded in the lexicon.

9. Special considerations for parts of speech (\ps)

There is a story about a baseball umpire who was asked, "How do you know which ones are balls and which ones are strikes?" He replied thoughtfully, "Well, some balls are clearly strikes, some balls are balls, and some are nothing until I call them." Problems in the categorization of parts of speech in the lexicon are like this.

There are frequently observed discrepancies between principles of linguistics and the practice of indicating parts of speech in the lexicon. The discussion here is aimed particularly at lexicographers in the early stages of compiling a dictionary. The general principles of determining parts of speech are not new, and are addressed to one degree or another in standard works on grammar or lexicography (Nida 1949, Zgusta 1971, Bartholomew and Schoenhals 1983, Givón 1984, 1990, Schachter 1985, Wierzbicka 1988).

We take it as a given that dictionaries are meant to be used and should therefore be user-centric (user-friendly) rather than compiler-centric (see §4.2). The way in which information is packaged in a dictionary must be adapted to the specific audience, but such adaptation must not compromise an accurate representation of the language. Regardless of which group of users is in focus, the information in a lexical entry on parts of speech (also referred to as 'word class' or 'form class'), in conjunction with the description of the grammar should enable the uninitiated user of a dictionary to understand—and hopefully to use—the lexeme in its appropriate syntactic contexts.

NOTE: *Parts of speech* in a lexicon is simply a tag that identifies the lexeme as a member of a *category* that shares a cluster of properties in its morphosyntactic network with other members of the same category. The parts of speech tag is a link between meaning and grammar.

The minimal information necessary to enable the dictionary user to make effective use of any part of speech category varies from language to language. The trouble is, however, very often such information is not in a dictionary, or it is misleading, or it is insufficient to be useful. We may get an approximation of the meaning, but the information on how the lexeme behaves, or how to use the lexeme is inadequate.

One could well ask whether we need parts of speech information in a dictionary at all, especially since such information seems to be of little interest or relevance to proficient speakers. We must recognize, however, that a dictionary is very often the first and most frequent resource a person (including linguists) consults when learning or studying a new language. It is for these 'outsiders' that information on parts of speech categories is most useful.

We also recognize as a secondary consideration the utility of the principle of transferability from one terminological system to another. For example, the cluster of properties for what is labeled as *noun* should overlap significantly with the cluster of properties that are generally labeled as *noun* cross-linguistically. For a group of local users, category labels should be adapted to the labels used for the national language where the associations with those categories are not in conflict. This facilitates a transfer to and from dictionaries of other languages, such as the national language or an international language.[1]

For bilingual dictionaries, the introduction must clarify whether the parts of speech categories reflect the source language or the target language. This is information is often missing, or often confused.

9.1 Common principles behind determining parts of speech

Using traditional parts of speech categories, and using the terms commonly accepted in the nation or region in which the language is spoken is certainly a place to start, but is something that must not be simply assumed or blindly accepted. In determining or refining parts of speech categories there is fairly broad acceptance of basic principles.

> **CAUTION**: Pinning linguistic labels on bits and pieces of a language is justifiable only where the structures of the language itself indicate contrastive patterns.

A fundamental principle underlying all analysis is determining whether two things are considered the *same* or *different* within the scope under scrutiny. An operating assumption is that it is preferable in a dictionary to associate similar forms that share a common thread of meaning. Parts of speech categories for a language are generally determined by comparing and contrasting the following criteria:

a) ***Form***: In some cases the structural form of an entire form class distinguishes it from other form classes. In Buru, prefixes can be distinguished from proclitics on the basis of form—prefixes always take the shape **e**C**–**, while proclitics can take any V and are of the shape CV (Grimes 1991:60). Also in Buru, certain classes of functors may be monosyllabic, but classes using content words (e.g. nouns and verbs) are never monosyllabic.

b) ***Function***: When we talk about an entire form class, or the behavior of a single lexeme in the syntax we usually refer to its 'function', or its range of functions—

[1]In SHOEBOX the **\ps** field is provided for English parts of speech, and **\pn** for the national language parts of speech. While the terminology between the two may be different (for example, **\ps n = \pn kb**), the categories should be the same, because one is targeting the categories of the vernacular, not of English or of the national language.

what it does, or how it (and things like it) behaves in different contexts. For example, in many languages that have prepositions, the function of the class of prepositions is to relate non-core arguments to the verb and to identify the semantic role that argument is playing. The function of prepositions contrasts with the function of verbs.[2] Schachter (1985:4) observes the preference "that the assignment of parts of speech classes is based on properties that are grammatical rather than semantic." Thus, defining nominals as the head of grammatical arguments in a clause is preferable to defining them as words that name persons, places, or things.[3]

c) ***Distribution***: The distributional behavior of a lexeme or a form class must also be taken into account. This includes the syntactic slot(s) it fills, as well as combinatory possibilities with affixes and with other form classes. In compiling a dictionary or writing a grammar, the well-attested phenomenon of *complementary distribution* is often overlooked in determining parts of speech categories relevant to a given language.

We refer to the combination of the above criteria as the 'morphosyntactic network' of a lexeme or a form class. In many languages assigning parts of speech to a lexeme is quite straightforward for the bulk of the lexicon—a noun is clearly a noun, a verb is a verb, and a preposition is a preposition. This chapter focuses on situations where categorization is not so straightforward.

9.2 Common areas of discrepancy between principle and practice

Assigning parts of speech in the lexicon is often problematic when there are discrepancies between principles and actual practice in lexicography.

a) Lexicographers may assign parts of speech on the basis of the gloss in the national (or international) language, rather than on the syntactic behavior of the form class in the language itself. In Buru, for example, we might be tempted to call **saa** an 'article' because it most commonly translates into English with the indefinite article 'a'. However, in exploring the whole morphosyntactic network it becomes clear that **saa** is a member of a closed class of what Grimes calls 'deictics' that share a variety of formal, functional, and distributional properties (C. Grimes 1991:167–175).

---

[2]Except, of course, in the case of prepositional verbs or serial verbs where a verb functions as a preposition might in another language.

[3]This characterization of a nominal is not tight enough for languages such as Tagalog, in which verbs are also used as clausal arguments (see Schachter 1985:9).

b) Lexicographers tend to remain committed to the parts of speech labels that they first assigned to a lexeme in the early analysis of a language (with associated assumptions about the behavior of that part of speech), even after those labels are shown to be inappropriate. Ideas about part of speech categories need to be refined and updated in the development of a lexicon to reflect developments in the understanding of the grammar.

c) Lexicographers generally assume 'word class' or 'part of speech' is inherent to the lexicon, and that every lexeme belongs fundamentally to a single part of speech category. Most lexicographers (and linguists) are not aware of operating on this assumption, but freely acknowledge it when it is brought to their attention. However, the empirical and theoretical basis for the assumption is problematic, and we discuss later the possibility that parts of speech for some parts of the lexicon may need to be defined syntactically, rather than lexically. After all, the whole notion of parts of speech is with reference to the syntax of a language.

d) When a lexeme can function in two or more classes (e.g. both nominally and verbally, or as a preposition and a conjunction), lexicographers tend to assume that it must be primarily one class, and only secondarily the other, assigning primacy on the basis of external (etic, rather than internal, emic) criteria. This is the 'flaw of the excluded middle'.

e) There is a tendency to assume certain word classes, such as 'adjective', are universal to all languages, and must therefore be in the language whose lexicon they are compiling.

f) Lexicographers often fail to distinguish verbal subcategories that are relevant to the language, assuming the only relevant primary division for verbs in all languages is limited to 'transitive' or 'intransitive'.[4] As described later in this chapter, the fundamental division for some types of languages is more complex than a simple binary distinction.

g) Lexicographers often tag multiple pronominal sets with terminology that is not appropriate to the type of language, such as using case terms (e.g. nominative-accusative or ergative-absolutive) for split-S languages or for pragmatically driven

[4]At a recent lecture a world-renowned linguist reiterated the notion that "all languages divide verbs into two types: transitive and intransitive." This simplification encourages linguists and lexicographers to be blinded to what distinctions languages actually *are* making where the fundamental divisions are more complex, such as in split-S languages, and blinded to notions such as 'ambitransitive' and 'intradirective'.

systems such as switch-reference systems.[5] In such languages labeling something as an 'ergative pronoun' or a 'nominative pronoun' reflects an inappropriate typology for the language.

9.3 Specific areas to watch out for

In the following sections we address various problem areas and suggest some ways in which parts of speech categories can more accurately reflect the language.

9.3.1 Views about the basis for assigning parts of speech

> **NOTE**: The traditional (and perhaps necessary) nature of a dictionary is as an artificial catalog of the lexicon, presenting a serial list of lexemes isolated from natural speech and organized around principles of retrievability of information.

That, together with ideas about what comprises a 'lexical entry' encourages linguists and lexicographers to slip into incorrect application of the Aristotelian principle that:

This lexeme cannot be both A and not A at the same time.

In other words, the thinking goes, this lexeme cannot be, for example, both a noun and a verb; therefore it must be primarily one and only secondarily the other (for example, through a zero-derivation),[6] or they must be two different lexemes. However, the problem arises out of the artificial nature of the dictionary in trying to assign parts of speech to lexemes in isolation. It is not the case in normal speech that a lexeme is functioning as both a noun and a verb *at the same time*. Where a lexeme is functioning in more than one category, it is either in different utterances, or in different syntactic slots within the same utterance. We explore below two areas in which the conflict commonly arises.

9.3.1.1 Are they adpositions or conjunctions?

A problem often occurs in assigning parts of speech to certain types of functors that operate in a variety of syntactic slots. For example, in English:

---

[5]This fallacy was reinforced at another lecture by a well-known linguist with the statement "75% of the world's languages are nominative-accusative and 25% are ergative-absolutive." This characterization blinds newcomers to well-documented language types such as split-S, active/non-active (~ stative-active), and Philippine-type languages which are numerically significant in the world's languages.

[6]This view often surfaces at linguistic seminars in lively debate over whether lexeme X is primarily category A or category B—and the implications for syntactic arguments that follow from that. The linear nature of a dictionary forces sense A to precede sense B, and it is part of the conventional culture of dictionary users to *assume* that the sense presented first is more basic—and for other reasons this makes good lexicographic sense.

(1)	He went ***to*** the store.	[relates verb to a non-core argument]
	He went ***to*** take a bath.	[relates verb to object complement purpose clause]

These are commonly handled in English dictionaries as separate lexemes (homonyms), yet they share the 'meaning' of energy directed toward a goal—one locative and the other purpose (see also Wierzbicka 1988). They are relating different types of syntactic units, and with the similarity in meaning they could be analyzed as the same lexeme with different functions in complementary distribution.

It is, in many cases, quite misleading to characterize functors of this sort merely as (or primarily as) a discourse particle, a clause-level conjunction, or an adposition (i.e. preposition or postposition). Many can function across a range of syntactic levels, linking constructions of varying scopes. The following contexts are from Buru (C. Grimes 1991:398).

(2)	PARAGRAPH$_1$. **Petu** PARAGRAPH$_2$	Linking paragraphs in a discourse
	SENTENCE$_1$. **Petu** SENTENCE$_2$	Linking sentences in a paragraph
	CLAUSE$_1$, **petu** CLAUSE$_2$	Linking clauses in a sentence (paratactic)
	Subject Verb **petu** CLAUSE$_2$	Subordinating a result clause (hypotactic)
(3)	**Tu dii**, DISCOURSE	'At that time,...' Introduces (cataphorically) the time setting in a discourse
	SENTENCE$_1$. **Tu** SENTENCE$_2$	Linking sentences in a paragraph
	CLAUSE$_1$, **tu** CLAUSE$_2$	Linking clauses in a sentence (paratactic)
	[N **tu** N]$_{Subject}$, Predicate	Coordinating nouns in an NP
	S – V – (O) – **tu** NP	Preposition

While some of these types of lexemes function exclusively as adpositions, some as conjunctions, and some as discourse particles, in a dictionary it is misleading to assign one of these classes to lexemes that can relate units of varying scopes. For this latter more flexible type, we prefer the broad term *relater*, rather than *preposition*, or *conjunction*. This issue of scope should be addressed in the grammatical introduction to a dictionary, but rarely is.

9.3.1.2 Are they nouns or verbs?

In many languages a portion of the lexicon is inherently and unambiguously nominal, while another portion is unambiguously verbal. But in many languages there is also a portion of the lexicon that may be either, according to its distribution and function in an utterance, such as the following examples from English.[7]

[7]The flexibility of this ambivalent portion of the lexicon may also vary between dialects of the same language. For example, Australian English can verbalize many words that are not able to be verbalized in American English. *She is flatting* (= *She is renting a flat/apartment*).

(4) She is going to ***sail*** around the world. [verbal]
He mended the ***sail***. [nominal]

(5) He went to ***photocopy*** the manuscript. [verbal]
He took the ***photocopy*** of the document away. [nominal]

(6) It looked like it was going to ***rain***. [verbal]
The ***rain*** got her wet. [nominal]

(7) He will ***shower*** under the tree. [verbal]
The ***shower*** is no longer working. [nominal]

(8) They found it hard to ***laugh***. [verbal]
They had a ***laugh***. [nominal]

Some lexicographers are tempted to argue etymologically for the primacy of membership in one form class over another, but unless there are clear synchronic derivational processes, the arguments may be much more difficult to substantiate and tend to appeal to elusive processes such as 'zero-derivation', which have no surface marking and which assume the primacy of one part of speech over another. Where 'zero-derivation' is warranted, there must be surface evidence somewhere in the morphosyntactic networks of the forms in question. Otherwise the claim of 'zero-derivation' is simply linguistic hocus-pocus.

Like English and other languages, Malay also has a number of lexemes whose function is distinguished only by its distribution within an utterance in an informal register,[8] such as:

(9) **Orang-nya** **jalan** **di** **jalan** **situ.**
person-ANAPH walk LOC path DIST.LOC
'The person went along that path.'

Such lexemes of ambivalent category membership are handled in different dictionaries variously as 1) different lexemes (homonyms), 2) the same lexeme in different distributions, 3) a compromise where they are viewed as separate but related lexemes (i.e. partial homonymy), or 4) by avoiding addressing the parts of speech issue altogether. Any four-year-old speaker of Malay knows the two are related, and not just because they sound the same. This kind of entry is handled in MDF as follows:

[8]Formal Malay would require derivational affixes such as **ber-jalan** for the verbal predicate use. One could argue on the basis of formal Malay that there is simply affix ellipsis for informal Malay. But this leaves at least two difficulties. First, what is the status of the unmarked base to which the verbal affixes (e.g. **ber-**, **meN--kan**) attach in the first place? Secondly, how can we argue for the elision of affixes that simply are not used in these contexts in informal Malay?

```
\lx jalan
\ps v
\ge go
\re go ; walk
\de go, walk
\ps n
\ge way
\re path ; trail ; road ; way
\de path, trail, road, way
```

jalan *v.* go, walk.
— *n.* path, trail, road, way.

There is extensive discussion in the literature on Austronesian languages of the Philippines and Taiwan (Formosan languages) as to whether the verbal construction should be interpreted as primarily nominal or verbal (see, for example, the discussion in Starosta, Pawley, and Reid 1982, and Ross, in press). The following derivations from the Paiwan root **kan** 'eat' are from Ferrell (1982:17, 106), adapted from Ross (in press).

(10)

Paiwan	Verbal construction	Nominal interpretation
k<əm>an	Actor pivot (neutral)	'eater' / 's.o. who eats'
kan-ən	Undergoer pivot (neutral)	'food' / 's.t. to be eaten'
k<in>an	Undergoer pivot (perfective)	'consumed food' / 's.t. eaten'
kan-an	Locative pivot (neutral)	'place where one eats'
si-kan	Instrumental pivot (neutral)	'eating utensil' / 's.t. to eat with'

CAUTION: The point is, the interpretation of these constructions as nominal or verbal depends largely on their distribution in the syntax. Unfortunately lexicographers do not tend to think in syntactically-oriented terms because of the lexically-oriented nature of dictionaries.

But as Wierzbicka (in press-a) observes about definitions, "words don't have any meaning in isolation, but only in sentences." And Halliday (1961:261) notes that, "a class is always defined with reference to the structure of the unit next above, and structure with reference to classes of the unit next below." In other words, word class—like meaning—is with reference to context.

9.3.1.3 Handling 'precategorials' (bound roots)

Many Austronesian languages have a number of lexical roots (content words) that are bound roots which never occur in an utterance without derivational morphology. For some bound roots, there is no internal evidence to say that one derived usage is more basic than another. Thus, one cannot, except by etic speculation, declare the root to be

primarily ‘nominal’, ‘verbal’, or whatever. Such bound roots, with reference to their form class membership are sometimes called *precategorials*.[9] For example in Buru:

(11)	**tea–**	‘(involving the planting of a post?)’
	tea-k	‘to jam s.t. postlike into the ground for use’
	tea-n	‘1) a (house)post’
		‘2) point of reference for kin group origins (place of original post?)’
	ep-tea	‘1) to live, stay, dwell (figurative from planting housepost?)’
		‘2) to sit down (extended sense from 1?)’
(12)	**mae–**	‘(involving a rigid object graspable in one hand)’
	mae-t	‘a fighting staff also used as a walking stick’
	mae-n	‘shaft (e.g. of spear); handle (e.g. of sword)’
	mae-k	‘to make a handle (e.g. of spear or sword)’
(13)	**bidu–**	‘(involving a cast-net)’
	bidu-k	‘to cast a cast-net’
	bidu-t	‘a cast-net’

In the last example above, one could argue either that 1) the nominal form uses /**–t**/ to derive the instrument that is characteristically used to perform the action of the verb, or 2) the verbal form uses /**–k**/ to derive a verb that is characteristically done using the noun. Both are legitimate explanations in the derivational paradigms of the language.

For an academic audience, precategorials can be handled as bound root lexemes (e.g. **mae–, tea–, bidu–**) with the surface derivations as subentries. But for a local audience this option is often not possible, since these bound root morphemes do not constitute a minimal possible utterance. For such an audience, one can work with the community to choose one derivation as the citation form [§5.4.4] with the other forms as minor senses, or else list each surface derivation as a separate lexeme. See §4.6 for extensive discussion with examples on how to organize lexical information in these two ways. One way of handling precategorials in MDF is as follows:

[9]Adelaar (1985:223) defines *precategorials* for his study of Proto-Malayic as “roots that do not occur in isolation, that is, roots which only occur in derivations and in compounds.” For Buru I expand the definition to include reduplication. E.g. **pani-n** ‘wing’, **p-e-pani** ‘HAVE wing (s.t. of which wing is the most salient feature)’, but never *[**pani-**] by itself. Some languages have a number of inherently reduplicated roots which never occur in the unreduplicated form. These roots could also be considered precategorials.

```
\lx bidu-
\ps Rt
\ge cast_net
\re *              [No reversal]
\de cast-net
\se bidut
\ps n
\ge cast_net
\re cast-net ; net (for
    casting)
\de cast-net
\se biduk
\ps vi
\ge cast_net
\re cast-net ; net (use by
    casting)
\de use a cast-net
```

bidu- *Rt.* cast-net.
bidut *n.* cast-net.
biduk *vi.* use a cast-net.

9.3.2 Verbal subclasses

For some languages more information is required than simply tagging verbal lexemes as *vi* (verb-intransitive) or *vt* (verb-transitive).

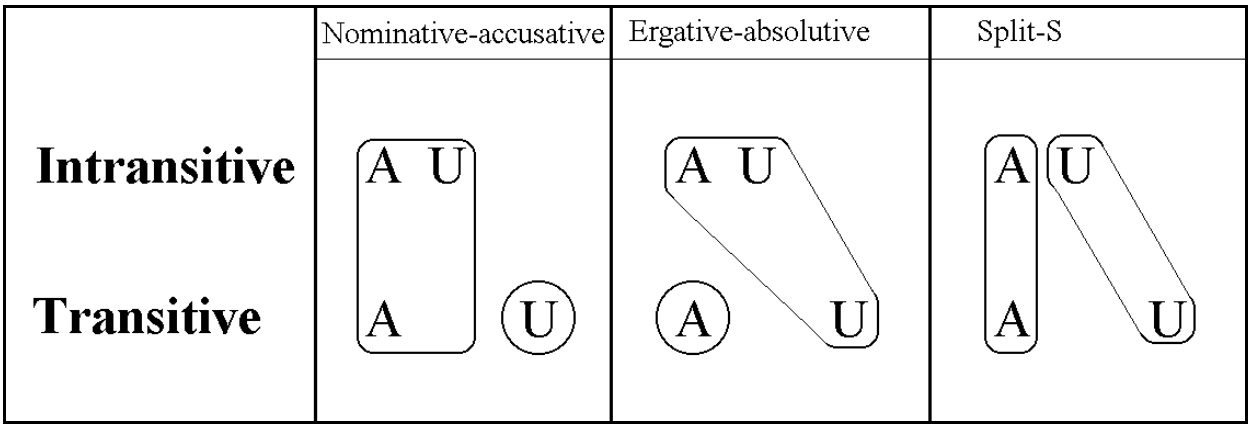
9.3.2.1 Split-S (split intransitive) languages

One type of language that requires a greater number of basic distinctions is split-S languages (Dixon 1979, 1994). In split-S languages the semantic Actor is encoded one way on both transitive and intransitive verbs and the semantic Undergoer is encoded a different way on both transitive and intransitive verbs. This pattern is called *split-S* (following Dixon 1979, 1994).[10] It is also known variously as split intransitive, stative-active, unergative-unaccusative, and in Government and Binding circles as 'ergative', with an unfortunate use of that latter term.

While there are a variety of types that fit into the split-S typology, perhaps the simplest is that in which the semantics of Actor and Undergoer are iconically mapped into the morphology or syntax of the language. If such a simple split-S typology is illustrated using the two English pronoun sets, it would operate something like that exemplified below. Active verbs (Actor-oriented) are DO or CAUSE type verbs (e.g. *do, make, go,*

[10]S in Dixon's system is the single argument of intransitive verbs. In a split-S system Actor and Undergoer are encoded differently on intransitive verbs—hence the name split-S. Dixon (1979) does not use Actor and Undergoer as primitives, but rather S, A, and O. We use the terms 'Actor' and 'Undergoer' in the sense of the macroroles described by Foley and Van Valin (1984).

hit, kill, break, return); non-active verbs (Undergoer-oriented) are BE or BECOME type verbs (e.g. *dark, ripe, white, sick, hungry, big, small, die, good, bad*).[11]



(14) Split-S system (S_A patterns with A; S_U patterns with O) [12]

he	**hit**	**him.**	[Active transitive;	A V O]
he	**ran.**		[Active intransitive;	S_A V]
	is sick	**him.**	[Non-active, postposed S;	V S_U]
him	**is sick.**		[Non-active, preposed S;	S_U V]

Split-S systems are fairly widespread within the Austronesian world, for example in Aceh, North Sumatra (Durie 1985), and in many languages in eastern Indonesia, such as Selaru, a language of southern Tanimbar (Coward 1990), and Dobel in the Aru Islands (J. Hughes, 1991). Buru is split-S in its verbal semantics, but shows an incipient switch-reference system in its pronominal typology (C. Grimes 1991).

All split-S languages must minimally distinguish *three* types of verbs in the lexicon, not just two, but dictionaries and wordlists published over the last century for split-S languages in eastern Indonesia have failed to do so. For Buru Grimes abbreviates the three types as *vt* (active transitive), *vi* (active intransitive), and *vn* (non-active verbs).

9.3.2.2 Intradirective or quasi-reflexive verbs

An additional verb type shows up in many Austronesian languages in eastern Indonesia and the Pacific. Active intransitive verbs tend to be verbs of motion or posture, such as

---

[11]The Selaru data and primary analysis are from Coward (1990). Some of the terminology and split-S framework reflect Grimes' adaptation of Coward's material. We avoid the label 'stative-active' that is widespread in the general literature for these types of languages, because the 'non-active' verbs are typically ambiguous in their internal aspectual interpretation as *imperfective* (process) or *perfective* (state). The label 'stative' at this macro level is thus highly misleading (see discussion in C. Grimes 1991:93-108).

[12]Relational grammarians call the S_A type verbs *unergative* and the S_U type verbs *unaccusative*. While there is nothing wrong with the terms for linguistic purposes, we do not recommend using these labels as parts of speech categories in a published dictionary as they severely limit the audience of effective users.

go, return, stand, sit, in which the person doing the action is also the one undergoing the action (their location or position is changed). For example, in *I go*, I am volitionally doing something that results in my location being changed. There is only *one* semantic referent, but some languages mark some (or all) active intransitive verbs of this sort *as morphologically transitive*. In some literature on Oceanic languages these are referred to as *intradirective*, or *reflexive* verbs (see Pawley 1973), and in other areas of the world as *quasi-reflexive* verbs.

(15) South Nuaulu (South-central Seram – R. Bolton 1990)

Ia	**pina**	**ona-te**	**i-sipu-i,**	**i-eu-i**
3s	female	big-NOM	3s-descend-3sU	3s-go-3sU

ria	**manahane.**
inland	outside

'The old woman got down and went outside.'

(16) Buru (archaic)

Kae	**oli-m**	**beka.**
2sA	return-2sU	first

'You should go home now.'

If these types of verbs contrast in a language with other active intransitive verbs that cannot be morphologically transitive, they must be indicated as a separate category in the lexicon, such as *vr* (verb reflexive).

9.3.2.3 Handling morphologically defined subclasses

It is still not always sufficient to identify a part of speech as, for example, *vt, vi,* or *vn*. Sometimes there are morphologically motivated subclasses within each category. For example in Buru, with a non-active verb (*vn*) we must also know whether it is an **em–** verb, an **eb–**verb, or a **–t** verb to know how it behaves in its morphosyntactic network.[13] MDF provides the **\pd** (paradigm) field to handle this additional information. Thus **\ps vn \pd –t** is minimal part of speech information for a non-active **-t** verb. Similarly, cataloging Buru active transitive verbs must distinguish whether they are **\pd –k** verbs or **\pd –h** verbs to know how they indicate pronominalized singular objects. Thus, **\ps vt \pd –h** is minimal information for a transitive **–h** verb in the lexicon. (See C. Grimes 1991:93ff.).

[13]Numerically indicated subclasses (e.g. Class I, Class II, Class III, etc.) seem to be very frustrating to everybody except the linguist who assigned those labels. An alternative such as the actual affixes that distinguish the subclasses in defined contexts broadens the audience of potential users (e.g. **em-**verbs, **eb-**verbs, etc.). This kind of morphological subclass is conventionalized in Spanish dictionaries in the citation form of verbs as **-ar, -ir,** or **-er** verbs.

9.3.2.4 Pragmatically motivated variants

Some languages, such as Fijian (Dixon 1988), have a clear morphologically motivated distinction between transitive and intransitive verbs or usages. Other languages have a group of verbs that are clearly and exclusively transitive, another group that are clearly intransitive, and a portion that may function in either capacity with no morphological distinction. The Buru data, for example, parallel the English:

(17)	**Da**	**ba**	**kaa**	**mangkau.**
	3s	DUR	eat	cassava
	'He is eating cassava.'			(NP object)

(18)	**Da**	**ba**	**kaa-h.**
	3s	DUR	eat-it
	'He is eating it.'		(pronominal object)

(19)	**Da**	**ba**	**kaa.**
	3s	DUR	eat
	'He is eating.'		(object suppression)

The above pattern of reducing the referential prominence of an argument through pronominalization and omission is a common strategy in discourse for languages that allow it (see Givón 1990). It is pragmatically motivated. The referential prominence of the object in example (19) above is completely reduced or suppressed through omission. Constructions like (19) occur commonly as predicate-focus constructions (action-prominence) where the referential identity of the object is unimportant or irrelevant, as in the situation: *Q: What is he doing now? A: He is eating [so don't bother him]*. But there is no morphological difference between the transitive use and the intransitive use of **kaa** 'eat', other than the presence or absence of the object.

Some linguists, perhaps motivated by the view that parts of speech are always and exclusively inherent to the lexicon, ignore syntactic and pragmatic issues, preferring to say **eat**$_1$ *vt* and **eat**$_2$ *vi* are two lexemes. We find this approach of (partial) homonymy highly unsatisfying. We prefer to say verbs like **eat** are included in that portion of the lexicon that is *ambitransitive* according to pragmatic issues, abbreviated as *vt/i*, or just *v*. The portion of ambitransitive verbs in English is fairly restricted, but in Buru most verbs that can take a syntactic object without morphological derivation are ambitransitive.

In languages where a portion of the vocabulary is ambitransitive in contrast with a portion that is obligatorily transitive, this contrast must be noted in the dictionary as a separate part of speech category.

9.3.3 Adjectives (versus nouns or verbs)

Some languages clearly have 'adjective' as a distinct part of speech, expressing such things as dimension, physical property, color, human propensity, age, value and speed. In some languages attributive modifiers in a noun phrase [NP] pattern closely with nouns, in other languages with verbs, and in others as a mixture (see Dixon 1982, 1991, Schachter 1985, Wierzbicka 1986 and in press). Buru, like many Austronesian languages, has no canonical (underived) class of 'adjective'—all attributive modifiers in an NP are derived from verb roots (both active and non-active).[14] For a few Austronesian languages in eastern Indonesia there does seem to be a closed class of a handful of underived adjectives (often in the form of inherently reduplicated roots), with the bulk of attributive modifiers in NPs being derived from verbs. With the exception of this small closed class of these true 'adjectives', there is often no morphological distinction between predicative and attributive uses of verbs.

(20) **Huma di em-kele.** [predicative]
house DIST STAT-tall
'That house is high.'

Da puna huma em-kele. [attributive]
3s do [house STAT-tall]$_{NP}$
'He made a pile house*.'* [Lit. a tall house]

Where there is a morphological distinction between predicative and attributive uses, it is clear that the attributive (i.e. adjectival) use is derived from the predicative use, not the other way around.

(21) **Da ba haa hede.** [predicative]
3s DUR big still
'He is still growing.'

Da puna huma haa-t. [attributive]
3s make [house big-NOM]$_{NP}$
'He is making a big house*.'*

(22) **Kau di beha**. [predicative]
wood DIST heavy
That wood is heavy.

Da wada kau beha-t. [attributive]
3s shoulder_carry [wood heavy-NOM]$_{NP}$
'He is carrying heavy wood *(on his shoulder).'*

[14]Nouns can also modify other nouns in Buru NPs, but behave quite differently from verb-derived modifiers in their morphosyntactic networks (C. Grimes 1991:178ff.).

(23) **Feten** **<u>boti</u>** **mohede.** [predicative]
millet white not_yet
'The millet isn't yet ripe.'

Da **ego** **labu-n** **<u>boti-t</u>.** [attributive]
3s get [shirt-GEN white-NOM]$_{NP}$
'She took the <u>white shirt</u>*.'*

To label what translates into an English adjective as 'adjective' for these languages fails to recognize the behavior of the lexemes as verbs in the greater morphosyntactic networks of the language.

9.4 Summary of \ps issues

Indicating parts of speech in the lexicon has been traditionally useful. The task of doing so is often straightforward and uncomplicated, but there are many potential pitfalls. The lexicographer must continually refine notions about parts of speech categories in a language and update the lexicon as understanding of the grammar increases. Parts of speech categories should be adequately defined to fit the language and make the dictionary a useful and productive tool.

9.5 Checking paradigms (\pd)

Some languages have obligatory indexing on the verb for one or more core arguments. For many Austronesian languages in the string of islands east of Bali, consonant-initial verb stems are not inflected for person and number of the subject, but vowel-initial stems are. In some languages, however, the paradigms are not complete for all possible combinations (also noted generally by Zgusta 1971:122). For example, in Tetun of central and east Timor (Therik and Grimes 1992), some verbs take the complete paradigm and others are only partial—the citation form of all these verbs is the *h*–form. The completeness of the paradigm can also vary across dialects.

Where there is inconsistency in the completeness or regularity of paradigms it is not economical to indicate the complete paradigm for every verb, but only for those that deviate from a norm. This information should be in the **\pd** or related fields. See §2.1.

	Complete Paradigms		Incomplete Paradigms		
PERSON	eat	bring	look	wait	pass by
1s	**kaa**	**kodi**	**karee**	**———**	**———**
2s	**maa**	**modi**	**maree**	**mein**	**mosi**
3s	**naa**	**nodi**	**naree**	**nein**	**nosi**
1px	**haa**	**hodi**	**haree**	**hein**	**hosi**
1pi	**haa**	**hodi**	**haree**	**hein**	**hosi**
2p	**haa**	**hodi**	**haree**	**hein**	**hosi**
3p	**raa**	**rodi**	**———**	**———**	**———**

9.6 Strategies for abbreviations

The increase in the last fifteen years of using interlinear examples to exemplify the use of linguistic data in its natural context has brought with it a bit of thinking about the advantages of certain strategies of abbreviation over others. Interlinear glossing has forced some of us to try and get rid of superfluous information and conventions in glossing and abbreviations that cause unnecessary spreading of the examples due to the length of the gloss.

(24) **ku-dengar-kan**
1s-hear -↑VAL

ku- dengar-kan
1SG-hear -↑VAL

ku- dengar-kan
1.PERS.SING-hear -↑VAL

One issue for abbreviations and glossing is choosing between *informal* and *formal* strategies. An informal strategy seeks to use the nearest translation equivalents in the target language (e.g. English) for both content words and functors. Thus, an Austronesian third singular genitive enclitic **–na** could be variously glossed as *-its, -his, -hers, -the*, and a free pronoun such as **aku** could be glossed as *I, me, my, mine*. A more formal strategy would seek to use consistent and possibly more technical terms for grammatical functions, such as *-3sG* or *1s*.

Simons and Versaw (1987:2–36) observe:

> The formal style uses abbreviations of technical terms for grammatical functions. These abbreviation are in all upper case letters and do not include a terminating period. For instance, one might use 'HAB' rather than 'always' as the gloss for a habitual verb aspect, or 'DEF' rather than 'the' in glossing a definite article. The use of upper case abbreviations to gloss functor morphemes is a fairly recent practice among linguists but it has gained widespread acceptance and can now be

considered a standard. There is, unfortunately, no definitive source of standard terms and abbreviations for text glossing. Rather the standard practice is for each investigator to devise his or her own set of abbreviations and to provide a complete listing of them in an introduction to the text collection or grammar sketch.

The style with upper case abbreviations for functor morphemes is in fact the standard for *Language*, the journal of the Linguistic Society of America (Bright 1984). It is advocated by Christian Lehmann (1982) in what is the only recent literature we are aware of on the subject of how to do interlinear text glossing. It is also evidenced in many recent textbooks (for instance, Comrie 1981, Foley and van Valin 1984, Givón 1984). (1987:2–76, 77).

The most complete listing of possible abbreviations we have found is in Lehmann (1982). This listing, which includes about 170 terms and proposed abbreviations, was compiled by collating terms and abbreviations from three published text collections. One of these, which provides a particularly good model, is Ronald Langacker's (1977–84) four volume set on Uto-Aztecan languages. Another good source for terms and abbreviations is numbers of the *Lingua Descriptive Studies* series. (1987:2–77)

One exception to the general rule of all upper case abbreviations in formal glossing is normally followed. This is in the glossing of pronouns. The convention for pronoun glosses combines a digit which designates the person and a lower case abbreviation for the number. For instance, '3sg' or '3s' for third person singular. (1987:2–36, 37)

Pronouns: Since most Austronesian languages have pronominal clitics of one or two letters (e.g. **ku–**, **mu–, –ng, –m, –n**/**–na**, etc.) it becomes important to use the shortest abbreviation possible for personal deixis that is not ambiguous. By using lower case for number we are then free to attach upper case grammatical or semantic tags. We suggest:

1s	(s = singular)	1sS	/ 1sSBJ	[subject]
2s		2sO	/ 2sOBJ	[object]
3s		3sG	/ 3sGEN	[genitive]
3sn	(non-human)			
1d	(d = dual)	1dP	/ 1dPOS	[possessive]
2d		2dH	/ 2dHON	[honorific]
3d		3dA	/ 3dACT	[actor]
1pi	(i = inclusive)	1piA	/ 1piABS	[absolutive]
1px	(x = exclusive)	1pxE	/ 1pxERG	[ergative]
2p	(p = plural)	2pD	/ 2pDAT	[dative]
3p		3pU	/ 3pUG	[undergoer]

Other functors: For grammatical categories it is good to use upper case with no period (full stop). A period is superfluous.

(25) DUR ‘durative’
HAB ‘habitual’
CAUS ‘causative’

For *portmanteau morphemes* (other than pronominal ones already taken care of like ‘*1sPOSS*’) it is common to use a period [.]. (Another convention encountered in the literature for portmanteau morphemes is the use of a colon [:], but a period is more common).

(26) PRES.PROG ‘present progressive’
PST.PRF ‘past perfect’

Hyphen [-] is, of course, a standard abbreviation for *morpheme breaks*. Among linguists in general there is inconsistent use of plus [+] and equals [=] in which the symbols sometimes appear to be there for principled reasons, and sometimes not. Plus [+] often indicates a grammatical clitic that is phonologically bound. Equals [=] is used to indicate reduplication of morphologically complex units (e.g. **ep-tilo=ep-tilo**).[15]

Standard non-technical abbreviations, of course, should be lower case. Some publishers do not allow them to be used at all: *etc., arch., cf., alt., e.g.*

Source languages for loans, if we abbreviate them at all, should be the minimum necessary for our purposes, using conventions widely accepted. E.g. *Eng.* rather than *Engl., Port.* rather than *Portug., Skt.* rather than *Sans.*

Indefinite terms should follow a single pattern. We suggest (although the periods take up extra space):

(27) s.t. ‘something’
s.o. ‘someone’
k.o. ‘kind of’

There will be some overlap for which choices have to be made. Some sort themselves out along the lines suggested above.

(28) gen. ‘generic (better to spell out)’
GEN ‘genitive’

(29) COMP ‘completed/completive/
complement/complementizer?’
CONT ‘continuative/contemplated/
contiguous?’

For a suggested starter list of abbreviations arranged alphabetically, see Appendix E.

[15]The use of equals [=] as the basic marker of a morpheme break, while used by some, tends to clutter the material visually and is not recommended.

9.7 RANGE SETS (consistency check for sets of abbreviations)

SHOEBOX allows the user to define master lists of abbreviations that SHOEBOX will check against. Thus, if the user compiles a master list of abbreviations for fields such as parts of speech (**\ps**), or semantic domains (**\sd**), or paradigms (**\pd**), then SHOEBOX can alert the user to misspelled or additional forms. New abbreviations can be added as needed, but the RANGE SETS feature of SHOEBOX provides a consistency check. Only forms actually used should be included. [See SHOEBOX manual for instructions on setting up the RANGE SETS feature].

10. Completing the dictionary

It is helpful near the beginning of a dictionary project to be aware of a number of tasks that will help facilitate the eventual completion.

10.1 Extracting topical subsets (e.g. kin terms, plant terms) from the master lexicon for analysis or for separate publication

While a good dictionary of a little described language can take 10–15 years to complete, there are often demands by sponsoring agencies, governments, local communities, and others to show that progress is being made along the way. Progress of this sort is most easily demonstrated by publishing and circulating *something*.

Some aim a preliminary publication as simply a dump of all the semi-edited work that has been completed in the lexical database to that point. These publications often have "A first dictionary of ...", "A preliminary dictionary of ...", "A concise dictionary of ...", "A shorter dictionary of ..." "A traveler's dictionary of ...", "A pocket dictionary of ..." or something similar in the title to indicate the incomplete nature of the work. These publications require a lot of special work that may or may not contribute directly toward the completion of the more complete dictionary.

> **TIP**: An alternative approach that we recommend is to work through different semantic domains in detail (e.g. kin terms, plants, cultivated plants, birds, fish), and to publish a series of separate volumes on each of these topical domains along the way toward publication of the complete dictionary. (See §6.4 for a discussion with examples of the **\sd, \th,** and **\is** fields.)

This alternative strategy allows the compilers to foster and incorporate community involvement along the way, and develop a community of readers who have a growing ability to use reference-type materials. Furthermore, these topically oriented volumes feed useful information to interested scholars, and demonstrate progress and competent work to government officials and sponsoring agencies—that is, if these are not also hasty dumps.

All primary work should be done in the main lexical database. If the information is flagged consistently, at the appropriate time one can extract the selected information into a separate database for processing through MDF by using the SHOEBOX FILTERS. For example, to extract kin terms and terms related to social structure (clan head, village head, etc.) one could use the following SHOEBOX FILTER (two examples are given—one simple and the other more complex):

```
\filt kin [sd|Nkin]
\filt kin [sd|Nkin] or [sd|Nsoc]
```

10.2 Writing an introduction to your dictionary

We recommend writing the first draft of your introduction after initial processing of the first 1,000 entries, adding refinements as you go along.

The basic purpose of the introduction to the dictionary is threefold:

1) To provide a brief orientation to the language and its speakers.

2) To provide a roadmap for using the dictionary.

3) To provide the information necessary for the dictionary to be usable as an independent (self-contained) volume. Use of the dictionary should not require the user to have a grammar of the language in one hand and an ethnography in the other.

Each topic covered in the introduction should be *relevant* to the dictionary and should be expressed *concisely*. Elaboration of the information found in the introduction to the dictionary should be included in a separate comprehensive grammar, an ethnography, and perhaps a history. The relative ordering of presentation of various issues should involve some creative thinking as to what information is more helpful.

If the dictionary is intended for publication in a linguistic journal, we recommend contacting the editorial board as to their formatting and organizational requirements. More specifically, the introduction should address the following:

1) Identify the primary audience and purpose for the dictionary. Also explain the overall organization of the dictionary information (e.g. give the ordering of the alphabet for the language). Give a total number of entries for the main dictionary and for the finderlist.

2) Briefly describe the *location* of the language, the number of people in the ethnic group, the *number of speakers*, and the regional context in which the language group is located.

3) Briefly describe any *historical events* (war, migrations, disease, colonization [European or otherwise], forced resettlement, intrusion of outside religions), or long-term activities (cross-ethnic marriages, general trade, coffee trade, slave trade, inter-tribal warfare, educational system(s)) that account for contact-induced language change and enable the reader to interpret the information in the **\et** (etymology) and **\bw** (borrowed word) fields.

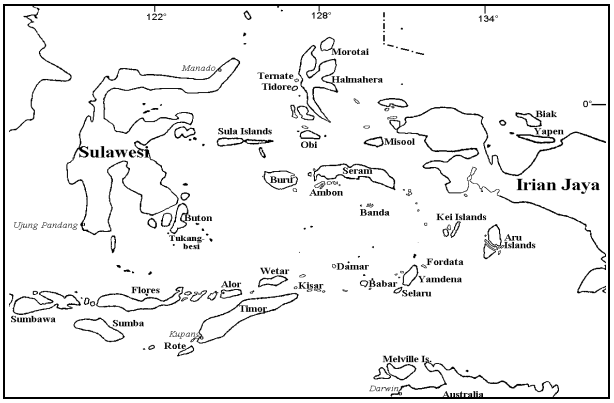
4) Provide a brief discussion about the language name and alternate names for the language if this is a relevant issue.

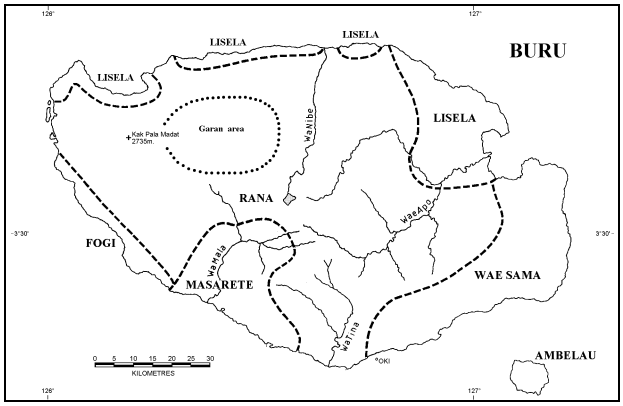
5) Mention the *linguistic classification* of the language (refer to the *Ethnologue*, B. F. Grimes 1992, or the more recent *Ethnologue language family index,* J. Grimes and B. F. Grimes 1993). Mention whether the classification is disputed. Mention *related languages* that might be known from the general literature and clarify how these are related. Avoid vague and relative terms like 'close' and 'distant'. Remember that some linguists will describe two unintelligible languages as 'close' that are less than 30% true cognate. Their framework and purposes may be different from yours.

6) List previously published works on the language.

7) Provide a brief *sociolinguistic profile*, including the *dialects*, the social registers, the patterns of lexical taboo, different speech patterns across genders or ages, or educated speaker usage, or whatever else will assist the users of the dictionary to get a dynamic view of the language and correctly interpret the **\ue** (usage), **\va** (variant), **\oe** (restrictions), **\lf** SynD (dialectal synonym), **\lf** SynR (register synonym), **\lf** SynT (taboo synonym), and **\lf** SynL (assimilated loan synonym) fields.

8) Provide *maps* in the introduction placing the language in its regional context, and a dialect map to help the reader understand the information on dialectal variants. It is surprising how many dictionaries of lesser known languages do not provide even a simple context map.





9) Provide a brief *phonology sketch*, a guide to pronunciation, and a guide to the *orthography* used in the dictionary, including a description of the morphophonemic processes that will enable the astute reader to approximately reconstruct the phonetics of polymorphemic forms from the information in a lexical entry. Explain all diacritics carefully. Supply a few well-chosen examples. Where there are

competing orthographies, you may need to provide a comparative table of equivalents to clarify the differences, with a brief word on why your particular orthography is used in the dictionary. (The reasons, may be linguistic, historical, political, social, etc.)

10) Provide a brief sketch of the *grammar* of the language, focusing particularly on how various *parts of speech* are defined and their distributional behavior. (See chapter 9). This section comprises the bulk of the introduction. For many users of a dictionary this is the section that can make it a good or a bad dictionary, a frustrating possession, or a useful resource. Remember to cover every part of speech referred to in the **\ps** field, and to give a few well-chosen examples to compare and contrast them with other, similar parts of speech. This is not the place to try and write a comprehensive grammar for an academic audience. That should be done in a separate volume. This is the place to summarize and illustrate key points discussed at greater length in the comprehensive grammar that are relevant for using the dictionary, whether the user be a layman or a professional linguist. An example of an introduction that is particularly complete in this area is Newell (1993).

11) Provide a brief *ethnographic sketch* to help the reader interpret entries on kinship, social structure, material culture, economics, agriculture, and cosmology. This should be concise, but useful. The fuller information should be found in a separate ethnography.

12) Provide a guide to *labels and abbreviations* used in the dictionary. Do not assume the reader is familiar with abbreviations that are conventionalized in the region or in the language family.

13) Provide a specific section describing *how to read a dictionary entry* in your dictionary. What information is presented first? What kinds of information are presented in an entry and what is the relative order of presentation? What do the different fonts represent (i.e. bold, italic, sans serif, etc.)? What is the structural hierarchy of an entry (subentries, senses, multiple parts of speech, etc.). How are homonyms marked, and how are homonyms cross-referenced? What do parentheses () mean? What do square brackets [] mean? What does an asterisk (*) mean? What do the different labels mean (*From: Etym: Usage: Ant: See: See main entry:* etc.)?

14) Provide a section describing how to use the *reversed finderlist*.

15) Provide a bibliography of all known references to the language, culture, and history of the language described in the dictionary. Include the sources used, for example, for flora and fauna.

10.3 Acknowledgments for the dictionary

The basic principle here is to be generous with your acknowledgments. Include those individuals who have invested their time in sitting down with the compilers and sharing their knowledge and insights. Mention community leaders, government officials, academics, consultants, and others who have had a role in the access, the process, or the production of the dictionary over the years since the initial field work. There may be organizations, such as private voluntary organizations, funding agencies, universities, government agencies, or others who have sponsored the field work or funded all or part of the effort. These should all be acknowledged graciously.

Once the dictionary is printed, make the effort and expense to ensure that key individuals and agencies, both local and national, receive a complimentary copy of the dictionary. This helps keep access to the region, the people, and the data open to yourself and to other researchers.

Appendix A: Alphabetized listing of field markers (with labels printed by MDF)

The following list is for reference purposes only. See §2.1 and other relevant sections for fuller explanation. [——— means 'none' or 'there is no label added for this field'; '...' means 'your text enclosed by'].

Field Codes	Function	English Label	National Language Label
\an	antonym	*Ant:*	*Lawan:*
\bb	bibliographical ref. for further reading	*Read:*	*Baca:*
\bw	borrowed word (loan)	*From:*	*Pinjaman:*
\ce	cross-reference gloss (English)		
\cf	cross-reference	*See:*	*Lihatlah:*
\cn	cross-reference gloss (national lang.)		
\cr	cross-reference gloss (regional lang.)		
\de	definition/explication (English)	———	
\dn	definition/explication (national lang.)		———
\dr	definition/explication (regional lang.)	*[Regnl: ...]*	*[Melayu: ...]*
\dt	date (entry last worked on)	———	
\dv	definition/explication (vernacular)	———	
\ec	etymology comment		
\ee	encyclopedic information (English)	———	
\eg	etymology gloss		
\en	encyclopedic info. (National lang.)		———
\er	encyclopedic info. (Regional lang.)		*[...]*
\es	etymology source		
\et	etymology (proto form)	*Etym:*	*Asal:*
\ev	encyclopedic info. (vernacular)	———	———
\ge	gloss (English)	(supplanted by \de)	
\gn	gloss (national language)		(supplanted by \dn)
\gr	gloss (regional language)	*[Regnl: ...]*	*[Melayu: ...]*
\gv	gloss (vernacular)	———	———
\hm	homonym/homophone	(subscripted)	
\is	index of semantics	*Semantics:*	*Kelompok:*
\lc	citation form (lexical citation)	———	
\le	gloss of **\lf** (English)		
\lf	lexical functions	(various, see §7)	
\ln	gloss of **\lf** (national language)		
\lr	gloss of **\lf** (regional language)		

Field Codes	Function	English Label	National Language Label
\lt	literally	*Lit:* '...'	*Lit:* '...'
\lx	***lexeme*** (headword/lemma)	———	
\mn	main entry form	*See main entry:*	*Lihatlah kata induk:*
\mr	morphology	*Morph:*	*Morf:*
\na	notes (anthropology)	*[Anth: ...]*	*[Antro: ...]*
\nd	notes (discourse)	*[Disc: ...]*	*[Wacana: ...]*
\ng	notes (grammar)	*[Gram: ...]*	*[Tata: ...]*
\np	notes (phonology)	*[Phon: ...]*	*[Fono: ...]*
\nq	notes (questions for investigation)	*[Ques: ...]*	*[Tanya: ...]*
\ns	notes (sociolinguistics)	*[Socio: ...]*	*[Sosio: ...]*
\nt	notes (general)	*[Note: ...]*	*[Cat: ...]*
\oe	only/restrictions (English)	*Restrict:*	
\on	only/restrictions (national language)		*Terbatas:*
\or	only/restrictions (regional language)		*[...]*
\ov	only/restrictions (vernacular)	*VerRestrict:*	*VerRestrict:*
\pc	picture [or graphic link]	(...)	
\pd	paradigm	*Prdm:*	*Pola:*
\ph	phonetic form (pronunciation)	[...]	
\pl	plural form	*Pl:*	*Jamak:*
\pn	part of speech (national language)		———
\ps	part of speech	———	
\rd	reduplication form(s)	*Redup:*	*Redup:*
\re	reversal (English)	**re:**[1]	**re:**
\rf	reference to written source (text or data notebook)	*Ref:*	———
\rn	reversal (national language)	**rn:**	**rn:**
\rr	reversal (regional language)	**rr:***[Regnl: ...]*	**rr:***[Melayu: ...]*
\sc	scientific name	*your text*	———
\sd	semantic domain	*SD:*	*Golongan:*
\se	subentry	———	
\sg	singular form	*Sg:*	*Tunggal:*
\sn	sense number	)	)
\so	source	*[Source: ...]*	*[Dari: ...]*
\st	status (for editing or printing)	*[Status: ...]*	———
\sy	synonym	*Syn:*	*Searti:*
\tb	table (chart)	———	

[1]The reverse fields and word-level gloss fields are not designed for printing, but these labels are given so that if the user wants to print these fields, they can be differentiated from the rest of the information in the entry.

Field Codes	Function	English Label	National Language Label
\th	thesaurus	*Thes:*	*Keluarga:*
\ue	usage (English)	*Usage:*	
\un	usage (national language)		*Kegunaan:*
\ur	usage (regional language)		*[...]*
\uv	usage (vernacular)	*VerUsage:*	*VerUsage:*
\va	variant forms	*Variant:*	*Bentuk lain:*
\ve	variant (English gloss or comment)	(...)	
\vn	variant (national language)		(...)
\vr	variant (regional language)		(...)
\we	word-level gloss (English)	**we:**	**we:**
\wn	word-level gloss (national language)	**wn:**	**wn:**
\wr	word-level gloss (regional language)	**wr:***[Regnl: ...]*	**wr:***[Melayu: ...]*
\xe	example (English free translation)	———	
\xg	example (gloss for interlinearizing)	******Not supported by MDF******	
\xn	example (national lang. free trans.)		———
\xr	example (regional lang. free trans.)		*[...]*
\xv	example (vernacular)	———	
\1d	first person dual inflection	*1d:*	*1d:*
\1e	first person plural exclusive	*1px:*	*1j:*
\1i	first person plural inclusive	*1pi:*	*1j:*
\1p	first person plural	*1p:*	*1j:*
\1s	first person singular	*1s:*	*1t:*
\2d	second person dual inflection	*2d:*	*2d:*
\2p	second person plural	*2p:*	*2j:*
\2s	second person singular	*2s:*	*2t:*
\3d	third person dual inflection	*3d:*	*3d:*
\3p	third person plural	*3p:*	*3j:*
\3s	third person singular	*3s:*	*3t:*
\4d	non-human or non-animate dual	*3dn:*	*3dn:*
\4p	non-human or non-animate plural	*3pn:*	*3jn:*
\4s	non-human or non-animate singular	*3sn:*	*3tn:*

(Nearly 100 field markers total)

Appendix B: Relative order of fields in an entry (with labels printed by MDF)

MDF reorders data fields to a consistent field order. This is made necessary by some of the formatting operations and has real advantage to the researcher in that minor inconsistencies in field order during data entry will not affect the consistency of the printed dictionary. The main disadvantage is that if you don't like the established order, you have to go inside the MDFDICT.CCT file and 'tweak' it. (This is not difficult for an experienced CC user, but not recommended for someone unfamiliar with it.) The following are listed in the basic order they are formatted by MDF. The exceptions are: 1) the **\lx, \hm** and **\lc** fields are flipped if the **\lc** field has data; 2) a gloss field (**\ge, \gn, \gr,** or **\gv**) does not print if there is a definition field counterpart (**\de, \dn, \dr,** or **\dv**); and 3) the reversal and word-level gloss fields are not intended to print; if you request them (through the Change Settings menu option), they are grouped together *after* the definition fields (not mixed in with them).

Your choice of audience when formatting begins determines which labels are used. For example, a triglot for a national audience will use national language labels. Regional language fields are not independent of the national language fields, so a diglot for the national language will include the regional language fields (unless you have altered the settings so that all regional language fields are ignored). At this point, you cannot specify a vernacular-regional dictionary. All regional language information is enclosed in square brackets ([ ]). [——— means 'none' or 'there is no label for this field'; '...' means 'your text enclosed by'].

Field Codes	Function	English Label	National Language Label
\lx	lexeme	———	
\hm	homonym number	(subscripted)	
\lc	lexical citation	———	
\ph	phonetic	[...] (only one for all languages)	
\se	subentry	———	
\ps	part of speech	———	
\pn	part of speech-national language		———
\sn	sense number	)	)
\gv	gloss-vernacular	———	
\dv	definition-vernacular	———	
\ge	gloss-English	(supplanted by a \de)	

Field Codes	Function	English Label	National Language Label
\re	reverse-English	**re:**[1]	**re:**
\we	word level gloss-English	**we:**	**we:**
\de	definition-English	_____	
\gn	gloss-national language	(supplanted by a \dn)	
\rn	reverse-national language	**rn:**	**rn:**
\wn	word level gloss-national language	**wn:**	**wn:**
\dn	definition-national language	_____	
\gr	gloss-regional lang. (with \gn)	*[Regnl:]*	*[Melayu:]*
\rr	reverse-regional lang. (with \rn)	**rr:***[Regnl:]*	**rr:***[Melayu:]*
\wr	word-level gloss-regional (with \wn)	**wr:***[Regnl:]*	**wr:***[Melayu:]*
\dr	definition-regional lang. (with \dn)	*[Regnl:]*	*[Melayu:]*
\lt	literal meaning	*Lit:* '...'	*Lit:* '...'
\sc	scientific name	(no label, but text as *underlined italics*)	
\rf	reference for example	*Ref:* (only one for all languages)	
\xv	example sentence-vernacular	_____	
\xe	example sentence-English	_____	
\xn	example sentence-national language		_____
\xr	example sent.-regional (with \xn)		*[...]*
\xg	example sentence-interlinear gloss	***(***not supported by MDF***)***	
\uv	usage-vernacular	*VerUsage:*	*VerUsage:*
\ue	usage-English	*Usage:*	
\un	usage-national language		*Kegunaan:*
\ur	usage-regional (combines with \un)		*[...]*
\ev	encyclopedic-vernacular	_____	_____
\ee	encyclopedic-English	_____	
\en	encyclopedic-national language		_____
\er	encyclopedic-regional language		*[ ]* (brackets only)
\ov	only (restrictions)-vernacular	*VerRestrict:*	*VerRestrict:*
\oe	only (restrictions)-English	*Restrict:*	
\on	only (restrictions)-national language		*Terbatas:*
\or	only (restrictions)-regional (with \on)		*[ ]*

[1]The reverse fields and word-level gloss fields are not designed for printing, but these labels are given so that if the user wants to print these fields, they can be differentiated from the rest of the information in the entry.

Field Codes	Function	English Label	National Language Label
\lf	lexical function	(\lf label, e.g. 'Spec', becomes the label)	
\le	lexical function-English	(combines with \lf)	
\ln	lexical function-national language	(combines with \lf)	
\lr	lexical function-regional language	(combines with \lf)	
\sy	synonym	*Syn:*	*Searti:*
\an	antonym	*Ant:*	*Lawan:*
\mr	morphemic representation	*Morph:*	*Morf:*
\cf	cross-reference	*See:*	*Lihatlah:*
\ce	cross-reference-English gloss	(combines with \cf)	
\cn	cross-reference-national gloss	(combines with \cf)	
\cr	cross-reference-regional gloss	(combines with \cf)	
\mn	main entry form	*See main entry:*	*Lihatlah kata induk:*
\va	variant form	*Variant:*	*Bentuk lain:*
\ve	variant comment-English	(...)	
\vn	variant comment-national language		(...)
\vr	variant comment-regional language		(...)
\bw	borrowed word	*From:*	*Pinjaman:*
\et	etymology	*Etym:*	*Asal:*
\eg	etymology-gloss	(combines with \et)	
\es	etymology-source	(combines with \et)	
\ec	etymology-comment	(combines with \et)	
\pd	paradigm	*Prdm:*	*Pola:*
\sg	singular form	*Sg:*	*Tunggal:*
\pl	plural form	*Pl:*	*Jamak:*
\rd	reduplication	*Redup:*	*Redup:*
\1s	1st person singular	*1s:*	*1t:*
\2s	2nd person singular	*2s:*	*2t:*
\3s	3rd person singular	*3s:*	*3t:*
\4s	singular non-human/non-animate	*3sn:*	*3tn:*
\1d	1st person dual	*1d:*	*1d:*
\2d	2nd person dual	*2d:*	*2d:*
\3d	3rd person dual	*3d:*	*3d:*
\4d	dual non-human/non-animate	*3dn:*	*3dn:*
\1p	1st person plural-general	*1p:*	*1j:*
\1e	1st person plural-exclusive	*1px:*	*1j:*
\1i	1st person plural-inclusive	*1pi:*	*1j:*
\2p	2nd person plural	*2p:*	*2j:*
\3p	3rd person plural	*3p:*	*3j:*
\4p	plural non-human/non-animate	*3pn:*	*3jn:*

\tb	table	———	
\sd	semantic domain	*SD:*	*Golongan:*
\is	index of semantics	*Semantics:*	*Kelompok:*
\th	thesaurus	*Thes:*	*Keluarga:*
\bb	bibliographic reference	*Read:*	*Baca:*
\pc	picture	(...) (parentheses, or a graphic link)	
\nt	notes-general	*[Note:]*	*[Cat:]*
\np	notes-phonology	*[Phon:]*	*[Fono:]*
\ng	notes-grammar	*[Gram:]*	*[Tata:]*
\nd	notes-discourse	*[Disc:]*	*[Wacana:]*
\na	notes-anthropology	*[Anth:]*	*[Antro:]*
\ns	notes-sociolinguistics	*[Socio:]*	*[Sosio:]*
\nq	notes-questions	*[Ques:]*	*[Tanya:]*
\so	source	*[Source:]*	*[Dari:]*
\st	status	*[Status:]* (only one for all languages)	
\dt	datestamp (a SHOEBOX field)	———	

(Nearly 100 field markers total)

Appendix C: Starter list of semantic domains (\sd)

Below is a suggested starter list of semantic domains. The list should be expanded and modified according to the structural and cultural constraints of the particular language being cataloged.[1] In using these categories one can be quite flexible in what is included under the label (e.g. nouns expressing emotions can also be included under Vemot), because the purpose of these things is grouping similar things together for analysis or separate publication.

Nagri	agriculture
Nanim	animal
Nboat	boat related
Nbody	body part
Ncult	material culture
Nfish	fish related
Nfood	food related
Ngovt	government
Nhouse	house related
Ninsect	insect
Ninstr	instrument
Nkin	kinship
Nloc	locative noun
Nnature	nature/meteorological
Npart	part of a larger whole
Nplant	plant
Nresult	noun of result
Nrit	ritual
Nsick	sickness/medicine
Nsocial	social relations (non-kin)
Ntime	time
Vaffect	affect (hit, kick, knock, hammer)
Vagri	agriculture
Vbody	bodily function
Vcarry	carry verb
Vcog	verb of cognition
Vcolor	color verb
Vcut	cutting verb

[1]For a detailed discussion of many of the verbal subtypes for English see Dixon (1991). His appendix (p. 363-369) includes a useful listing of examples of the subtypes.

Veffect	verb of effect
Vemot	verb expressing emotion
Vevent	verb naming or characterizing a whole event
Vexchange	verb of exchange (give, receive, take, get)
Vhit	hitting verb
Vhold	holding verb
Vhunt	hunting related
Vmotion	verb of locomotion
Vposture	verb of posture or rest
Vrit	verb describing ritual
Vsee	verb of perception
Vsize	verb of dimension
Vsocial	verb expressing social relationship
Vspeak	speech-act verb
Vspeed	verb of speed
Vtouch	touching verb
Vvalue	verb expressing value
Vweath	weather verbs (rain, fog)
Vweight	verb expressing weight

ADJage	age
ADJbodily	bodily function
ADJcol	color adjective
ADJemot	emotion/human propensity
ADJphys	physical property (hard, clean, hot)
ADJsize	size/dimension
ADJspeed	speed
ADJtext	texture
ADJval	value (good, bad, nice)

See cautions about distinguishing between verbs and adjectives in chapter 9. See Dixon (1991) for more ideas.

Variations of the above information can be chosen according to the aesthetics of the compiler. Some alternate possibilities are as follows:

Option 1	Option 2	Option 3
Nagri	nAgri	Agriculture
Nbody	nBody	Body
Vcarry	vCarry	Carry
Vcut	vCut	Cut
ADJsize	adjSize	Size
ADJspeed	adjSpeed	Speed

Appendix D: Alphabetized starter list of lexical functions

This present list is intended only to help people get started and help them with the bulk of what they will find. Those who want to become proficient users of additional lexical functions, including the use of composite functions (e.g. CausIncep of *dark* = darken [transitive]; IncepN_0 of *storm* = break) are referred to J. Grimes (1987).

Ant	Antonym
Caus	Causal
Compound	Lexicalized compound using headword not easily handled by other lexical functions
Cpart	Counterpart (complement, conversive)
Degrad	Degraded degree or state
Feel	Feeling or sensation associated with headword
Gen	Generic
Group	Collective/group
Head	Head or leader of group
Idiom	Idiom
Mat	Material used to make headword
Max	Superlative degree of headword
Min	Diminished degree of headword
Nact	Actor noun
Nben	Benefactee noun
Ndev	Deverbal noun
Ninst	Instrumental noun
Ngoal	Goal of action
Nloc	Locative noun
Nug	Undergoer noun
ParS	Parallelism representing *Same* as headword
ParD	Parallelism representing *Different* end of scale
Part	Part of headword
Phase	Phase of headword
Prep	Preparatory activity
Res	Consequence or resulting state
Serial	Conventionalized serial verb combination not clearly handled by other lexical functions
Sim	Similar type at same level of hierarchy
Sit	Situation or activity typically associated with headword
Sound	Sound associated with headword
Spec	Specific (kind of, type of, species)
Start	Beginning phase of headword (inceptive)

Stop	Final phase of headword (cessative)
Syn	Synonym (same range of meaning)
SynD	Synonym in another dialect of the same language
SynL	Loan synonym fully assimilated into language
SynR	Synonym in another register of same language
SynT	Taboo synonym
Unit	Single occurrence of headword
Vwhole	Verb of the whole
Whole	Whole of which the headword is a part

Appendix E: Starter list of abbreviations

The principles behind certain strategies for abbreviations are discussed in §9.6. Below is a suggested starter set of abbreviations for parts of speech and interlinear glosses.[1] Where several forms compete for the same abbreviation (e.g. P-patient, P-possessive, P-parent), we suggest selecting the short form for either the most frequent abbreviation or the shortest vernacular morpheme.

Parts of speech:

ADJ	Adjective		
ADJR	Adjectivizer	MDL	Modal
ADV	Adverb		
ADVR	Adverbializer	NEG	Negative
AFFM	Affirmative	NEGimp	Negative imperative
AL	Alienable	NOM	Nominative
AN	Animate	NOMR	Nominalizer
APPL[2]	Applicative	n	Noun
ART	Article	NUM	Number
ASP	Aspect		
AUX	Auxiliary	PTCL	Particle
		PART	Participle
CLASS	Classifier	PAUS	Pause word
CMPAR	Comparative	PL	Plural
CMPLR	Complementizer	POSS/P	Possessive
CNJ	Conjunction	POSSR	Possessor
COND	Conditional	POST	Postposition
CONF	Confirmative	PREP	Preposition
CONN	Connective	PRO	Pronoun/pronominal
COP	Copula	PropN	Proper noun
DECL	Declarative	Q	Query/Question/Interrogative
DEIC	Deictic (spatial & temp.)	QNT	Quantifier
DEM	Demonstrative		
DIR	Directional	REC	Reciprocal
		REL	Relative(izer)
EVID	Evidential	RFLX	Reflexive
EXASP	Exasperative	RLR	Relater
EXIST	Existential		

[1]For an alternative list and framework for organizing lexical data, see the SHOEBOX manual.

[2]Not a brand of computer.

		TAM	Tense-Aspect-Mood
FOC	Focus marker	TIME	Time expression
		TNS	Tense
HORT	Hortative	TR	Transitive(izer)
ID	Idiom	v	Verb/verbal
IMP	Imperative	vi	Intransitive verb
INTJ	Interjection	vm	Middle verb (non-agentive passive)
INT/Q	Interrogative		
ITR	Intransitive(izer)	vn	Non-active verb
		vp	Passive verb (agentive)
		vr	Reflexive/quasi-reflexive /intradirective
LIG	Ligature	vt	Transitive verb
LOC	Locative	vt/i	Ambitransitive verb

General glosses and abbreviations:

A	Actor	HON/H	Honorific
ABL	Ablative	HUM	Human
ABS	Absolutive	i.e.	that is
ACC	Accusative	IMM	Immediate
ACMP	Accompany[3]	IMPRF	Imperfective
ACT	Active/Actor	IMPRS	Impersonal
ADDR	Address	INAL	Inalienable
ADVNC	Advancement (IO → DO)	INAN	Inanimate
ADVS	Adversative	i/INC	Inclusive (1pi)
AFFT	Affective	INCEP	Inceptive
AG	Agent/agentive	INCHO	Inchoative
ALL	Allative	INDEF	Indefinite
AN	Animate	INF	Infinitive
ANTP	Antipassive	INST	Instrumental
arch.	Archaic	IO	Indirect Object
ATTR	Attributive	IRR	Irrealis
		IT	Iterative
BEN	Benefactive		
		JUSS	Jussive
CAUS	Causative		
CESS	Cessative	k.o.	kind of
CIRC	Circumstantial		

[3]Same as COM (Comitative).

COLL	Collective
COM	Comitative
COMP	Completive
CONC	Concessive
CONT	Continuative
DAT	Dative
DEF	Definite
DER	Derivational
DES	Desiderative
DIM	Diminutive
DIST	Distal
DISTB	Distributive
DO	Direct Object
DUB	Dubitative
DS	Different Subject
DUR	Durative
e.g.	for example
EMPH	Emphatic
ERG	Ergative
etc.	etcetera
e/EXC	Exclusive (1pe)
EXCLM	Exclamatory
FACT	Factitive
F/fem.	Feminine (3sF)
FIG	Figurative
FREQ	Frequentative
FUT	Future
GEN/G	Genitive (1sG)
GER	Gerund(ive)
HAB	Habitual
RES	Resultative
sp.	Species
spp.	Species (plural)
s.o.	Someone

Lit.	Literally
MAN	Manner
M/masc.	Masculine (1sM)
MOD	Modifier
NARR	Narrative
NEC	Necessity
NFUT	Non-future
NHUM	Non-human
O/OBJ	Object (3sO)
OBL	Oblique
obs.	Obsolete
opp.	Opposite
OPT	Optative
PAT/P	Patient
PTT	Partitive
PASS	Passive
PAST	Past
PRF	Perfective
PERS	Personal
PIV	Pivot
PRES	Present
PROG	Progressive
PROX	Proximal
PURP	Purpose
QUOT	Quotative
REAL/R	Realis
RED	Reduplication
REF	Referential/Term of reference
REM	Remote
REP	Repetitive
TEMP	Temporal
TOP	Topic
TOPR	Topicalizer
U / UG	Undergoer

s.t.	Something		
S/SUBJ	Subject (2sS)	viz.	namely
SPEC	Specific	VOC	Vocative
SS	Same subject	VOL	Volitional
STAT	Stative	VP	Verb Phrase
SBJV	Subjunctive	vs.	versus
SUP	Superlative		

Kinship:

B	brother	M	mother
C	child	(m.s.)	male speaking
D	daughter	P	parent
e	elder	S	son
F	father	W	wife
(f.s.)	female speaking	y	younger
H	husband	Z	sister

[This system allows combinations such as WBW 'wife's brother's wife', MB 'mother's brother', eB(f.s.) 'elder brother (female speaking). These abbreviations are useful for short interlinear glosses.]

Loan sources:

AM	Ambonese Malay	Jav.	Javanese
Ar.	Arabic	KM	Kupang Malay
Bug.	Bugis	Mak.	Makassar
Btn.	Butonese (generic)	Mly	Malay
Du.	Dutch	Port.	Portuguese
Eng.	English	Skt.	Sanskrit
Fr.	French	SM	Standard Malay
Ger.	German	Sp.	Spanish
Ind.	Indonesian	Sw.	Swahili
Jap.	Japanese	TM	Ternate Malay

Conventions:

*	Reconstructed form (historical)
**	Intermediate hypothetical form (historical)
[...]	Implicit information [square brackets]
/	Optional interpretation [or]
-	Morpheme boundary
.	Portmanteau morphemes (PRES.PROG)
=	Reduplication of complex units
~	Varies with

Appendix F: Enhancements and changes from v0.9 and v0.95

F.1 Enhancements in MDF v1.0

The most exciting aspect of MDF 1.0 is the automatic formatting options that are now available. Older versions supported only triglot or diglot with the national language output; no vernacular-English diglot was available. Now, you can choose the audience (English or national language) and whether you want the output to be triglot or diglot. The choice of audience then determines which diglot is produced.

A related enhancement is the ability to tell MDF to not format your example sentences (ignoring **\rf, \xv, \xe, \xn, \xr,** and **\xg** fields). This is useful for producing drafts, or when your example sentences are in need of serious work but you need a hard copy of the rest today for someone else.

You can also tell MDF to print or ignore your notes fields (ignoring the **\nt, \np, \ng, \nd, \na, \ns,** and **\nq** fields).

These options save the user from having to go through the Change Settings option and mark each of these fields as "pitched (disabled)" fields every time he or she wants to print out a variant dictionary format. Thus, printing a national language diglot with no notes and an English diglot with everything, is now a matter of answering two questions.

Taken together (triglot, diglot, national language or English audience, examples, and notes), these simple choices can produce 16 different dictionary formats (hopefully this is enough to meet the needs of most people). But if not, you still have the ability to set up MDF to ignore certain fields through the Change Settings menu option.

The Change Settings option now requests the name of the vernacular language and the name of the national language. These are stored and used later for formatting the dictionary and finderlists. This saves the user from having to answer these questions for each type of output. If you wish to change these names, simply select the Change Settings option again. (Selecting the Change Settings option again will *not affect* the fields you have already selected to be included or discarded. Choosing the Reset option will revert the fields back to the default settings and will erase the language names.) If you do not want to bother with "settings" then don't. MDF will ask for the language names as needed.

F.2 Changes from MDF v0.9 and 0.95

The following addresses the field marker and character formatting changes that have been implemented in this new version of MDF. To make these changes to your lexical database, you can use the change table UPDATE.CCT supplied on the release disk. (Be

sure to copy your original lexical database to floppies for safekeeping *before* going any further.)

> **CAUTION**: The CC table, UPDATE.CCT, assumes that your original lexical database followed the guidelines included with the 0.9x versions of MDF. *DO NOT use this CC table on your database if it does not conform to the older 0.9x standards!*

To use UPDATE.CCT type **CC** at the DOS prompt (what you type is bold):

C:\MDF>**cc**<ENTER>

The Consistent Changes program will display:

```
Consistent Change 7.4, 15–May–90 Copyright 1987–1990 SIL Inc.
Changes File? update.cct
Output File?  newlex.db
Input File?   lexicon.db                    (if that is the original name)
Next input file (<RETURN> if no more)?      (Press the <ENTER> key)
```

When you are asked for the input filename, give the name (and path, if needed) of your lexical database. CC will not alter your lexical database in any way. Just be sure *you don't give the original lexical database name as the **output** filename*. You can destroy your data that way!

The output file should now be just like your original database, except that it has the updated field markers and the new character formatting codes. But do not delete your original lexical database until you are sure that the new file is accurate (a directory listing should show the new file somewhat larger than the original). Also, be sure to tell SHOEBOX of the new filename.

F.2.1 Changes in field markers

The main changes between 0.9x and 1.0 involve the more generalized language references (which also involved a program name change from 'Maluku Dictionary Formatter' to 'Multi-Dictionary Formatter'). Basically, 'Indonesian' is now 'national language', and 'Malay' is now 'regional language' in all the documentation and field marker codes (e.g. **\gi** 'gloss Indonesian' has been changed to **\gn** for 'gloss national language'). These system changes in field markers required some shifting around of other codes as well to fit the common paradigm. For example, the original **\le** marker (for 'lexical entry' or 'lexeme') was now needed for the 'English gloss of a lexical relation field,' so the old **\le** has been changed to **\lx** for 'lexeme'.

The existing structure has also been embellished (with enhanced lexical functions, cross-references, etymology, variants, restrictions, and encyclopedic information). One field (**\vg**) was discontinued.

The following is a table depicting most of the changes that have been implemented in MDF v.1.0. On the left are the old field markers while on the right are the replacements.

'\le'	>	'\lx'	(lexical entry or lexeme)
'\gi'	>	'\gn'	(gloss—national)
'\ri'	>	'\rn'	(reverse gloss—national)
'\wi'	>	'\wn'	(word gloss—national)
'\di'	>	'\dn'	(definition—national)
'\xi'	>	'\xn'	(example sentence translation—national)
'\ui'	>	'\un'	(usage—national)
'\gm'	>	'\gr'	(gloss—regional)
'\rm'	>	'\rr'	(reverse gloss—regional)
'\wm'	>	'\wr'	(word gloss—regional)
'\dm'	>	'\dr'	(definition—regional)
'\xm'	>	'\xr'	(example sentence translation—regional)
'\um'	>	'\ur'	(usage—regional)

For the sake of consistency **\en** 'ethnographic notes' has been combined with **\na** 'notes—anthropology', and **\sl** 'sociolinguistic notes' has been renamed to **\ns** 'notes—sociolinguistics'.

'\en'	>	'\na'	(notes—anthropology)
'\sl'	>	'\ns'	(notes—sociolinguistics)

The earlier versions of MDF gave only one field marker for lexical relations (**\lr**). This was recognized as inadequate, but at earlier stages of MDF development it was unclear as to how people were encoding lexical relations on the computer. Grimes has documented how he and others are using this system (see chapter 7 and C. Grimes 1987, 1994), and has suggested the following field codes:

'\lf'	(lexical function)
'\le'	(lexical function gloss—English)
'\ln'	(lexical function gloss—national language)
'\lr'	(lexical function gloss—regional language)

The term 'lexical relations' was changed to 'lexical functions' to align it with the wider literature and to allow **\lr** to consistently refer to regional glossing. Note that **\le** (the old KEY field marker) is now used for English glossing of **\lf**. Be sure to convert all of your old key field markers to **\lx** before implementing this feature. (If you use UPDATE.CCT to convert your database, this is handled for you.)

The following gives an example of how lexical function field bundles are used.

\lf Syn = asumwany2
\le high water mark

\ln air pasang
\lr

MDF combines the gloss fields into the **\lf** field and formats them so they will print as follows:

Syn: **asumwany**$_2$ 'high water mark' '*air pasang*'.

There can be multiple groups of these field bundles:

\lf Syn=mlay
\le true
\lf Ant = sal
\le wrong, false

which will be separated with a semicolon:

Syn: **mlay** 'true'; *Ant:* **sal** 'wrong, false'.

The markers **\sy** 'synonym' and **\an** 'antonym' are still supported for those who wish to encode these lexical relations directly without using the **\lf** field bundles. But **\sy** and **\an** do not support glossing; they only allow for the vernacular cross-reference to be given. The conversion table UPDATE.CCT automatically converts all **\sy** and **\an** fields to **\lf Syn =** and **\lf Ant =** fields and then inserts a blank **\le** and **\ln** field for each **\lf** field. (It is assumed that most will not be using the **\lr** field, but it is available for those who need it.) For example:

\sy mlay

becomes

\lf Syn = mlay
\le
\ln

The user can then go through and fill in the **\le** and **\ln** fields at a later time (or leave them blank if preferable).

By analogy the **\cr** 'cross-reference' field has been converted to the following field bundle:

'\cf'	(cross-reference)
'\ce'	(cross-reference gloss—English)
'\cn'	(cross-reference gloss—national language)
'\cr'	(cross-reference gloss—regional language)

There can be more than one bundle per entry, subentry, or sense. (Note that the bundles need not use all of the fields.) UPDATE.CCT inserts a blank **\ce** and **\cn** field for every reference in an old **\cr** field. For example:

```
\cr -kw, -mw, -na
```

becomes

```
\cf -kw
\ce
\cn
\cf -mw
\ce
\cn
\cf -na
\ce
\cn
```

After you fill in the **\ce** fields,

```
\cf -kw
\ce my
\cn
\cf -mw
\ce your
\cn
\cf -na
\ce his
\cn
```

this prints out as:

See: **-kw** ‘my’; **-mw** ‘your’; **-na** ‘his’.

If you left the **\ce** fields blank, it would print out as:

See: **-kw**; **-mw**; **-na**.

(This is about what you would have gotten with the old method.)

This glossing capability will enhance the usefulness of the printed dictionary, since it will give the user an idea of what a reference means without having to actually flip over to that entry.

The use of ‘etymology’ in the old MDF documentation was weak. It really addressed loan or borrowed words rather than proto forms (which is what one would expect **\et** to refer to). So the old **\et** has become **\bw** ‘borrowed word’.

‘\et’ > ‘\bw’(borrowed word)

The **\pf** ‘proto form’ field has been changed to **\et** ‘etymology’ (this is a more accurate use of the terminology). If you convert your database on your own (using macros, etc.) be sure to convert all original **\et** fields to **\bw** fields *before* you convert **\pf** fields to **\et** fields. If you use UPDATE.CCT, this will not be a problem.

Like the **\lf** and **\cf** fields the new **\et** field supports a type of bundling.

‘\et’	(etymology)
‘\eg’	(etymology—gloss)
‘\es’	(etymology—source)
‘\ec’	(etymology—comment)

For example:

\et *tebel
\eg thick (dimension)
\es PANDW
\ec metathesis?

By default, this bundle will print out as:

Etym: *tebel ‘thick (dimension)’.

But if you request to include the **\es** and **\ec** fields through the Change Settings menu option, it will print out as:

Etym: *tebel ‘thick (dimension)’ PANDW (metathesis?).

Do not forget to include the ‘*’ in the **\et** field. Also, UPDATE.CCT will insert a blank gloss (**\eg**) field for each old **\pf** it converts to **\et**.

MDF will now format the **\ph** ‘phonetic’ field with square brackets, so that:

\ph apa

will print as:

[apa]

The font associated with the data in the **\ph** field is determined by the PH style in the MDFDICT.STY stylesheet. So, by changing the stylesheet, you can use a phonetic font for

this field. (The square brackets are not included in this PH style—they are formatted with the standard font.) **\ph** can be used in relation to both **\lx** (lexeme) and **\se** (subentry).

We have added encyclopedic fields for those who want their lexicon to be more of a cultural knowledge base. These fields are:

'\ev'	(encyclopedic—vernacular)
'\ee'	(encyclopedic—English)
'\en'	(encyclopedic—national)
'\er'	(encyclopedic—regional)

These are printed with no label (though the regional language field will be bracketed with square brackets).

The 'Usage' fields (**\ue, \un,** and **\ur**) now have a vernacular counterpart, **\uv**, for monolingual dictionaries. The vernacular field is labeled as '*VerUsage:*'

'Only' fields (**\ov, \oe, \on,** and **\or**) have been added to denote semantic or grammatical restrictions pertinent to the headword. This field is given the label '*Restrict:*'

A **\mr** 'morphemic representation' field has been added to provide a morpheme-by-morpheme breakdown of polymorphemic lexemes. This field is given the label '*Morph:*'

A **\lt** 'literal' field has been added for clarifying the literal meaning of idioms, etc. This field is given the label '*Lit:*' It also adds single quotes around the meaning.

A **\bb** 'bibliography' field has been added for recording bibliographical references to where the lexeme is treated at greater length (grammatically or ethnographically). This field is given the label '*Read:*'

A **\pn** 'part of speech—national' field has been added to allow for specifying the part of speech using labels found in national language dictionaries. MDF requires that the **\pn** field follow the **\ps** field:

\ps n	(noun)
\pn kb	(the national abbreviation for 'noun')

If the order is reversed, MDF will not function properly. MDF will format the **\pn** field only if you specify that the output is for a national audience. When a national audience is specified, the **\pn** field will replace the **\ps** field. But if there is no **\pn** field or if it is empty, the **\ps** field will be output for the national audience as for an English audience.

In the conjugation form fields, the glaring oversight of not including first-person inclusive and exclusive fields is corrected. These are **\1i** and **\1e**, respectively. The field marker **\1p** is still retained for those who work with languages that do not make this distinction. Also, 'dual' verb forms are now supported with **\1d, \2d, \3d, \4d** (non-animate, non-human).

The **\vg** 'vulgar' field is no longer supported (it didn't work right, and it was too limited in function). We are suggesting that the **\ue** 'usage—English' or **\st** 'status' fields could be used for encoding this type of information. UPDATE.CCT converts the '**\vg**' field to '**\ue** Vulgar'. If you wish to discard any vulgar entry, subentry, or sense from a printed copy, first format the dictionary normally, and then use SEARCH (or EDIT FIND) to locate 'Vulgar'. This will allow you to delete them out of the final copy. (You will be able to do this more accurately than with the old MDF program.)

F.2.2 Changes in character formatting codes from v0.9x

Language *font* codes are now mnemonically 'font...' rather than 'language...' In other words, the font code for English is 'fe:' (rather than 'le:'). Also as with the field codes, 'Indonesian' is now 'national', and 'Malay' is now 'regional':

'fv:'	for vernacular	(from 'lv:')
'fe:'	for English	(from 'le:')
'fn:'	for national	(from 'li:')
'fr:'	for regional	(from 'lm:')

We have also added 'standard,' bold, and italic fonts as well:

'fs:'	for standard font
'fb:'	for bold font
'fi:'	for italic font

These fonts are supported as character styles in the stylesheet, so they can be modified at any time. The standard font is used in MDF for formatting most information fields (**\rf, \lt, \pd, \lf, \is, \th, \sd, \bw, \et,** and **\cf**), as well as for punctuation. The labels used in MDF to mark the different fields (like the '*See:*' for the cross-reference field) are all encoded with the FL style (mnemonic for 'font—label'). With this style, you can change all labels in your dictionary to a different point size or font in one quick step.

Specifying underlined characters is now:

'uc:'	for underlined character	(from 'un:')
'ui:'	for underlined italic	(from 'us:')

Appendix G: Files and programs used by MDF

G.1 Print tables, etc. used by MDF

README	DOC	
MDF	DOC	(on-line Overview)
MDF	STY	(for Overview)
MDF	BAT	(the MDF program)
MDF	ICO	(an icon you can use in Windows)
MDF1	ICO	(an icon you can use in Windows)
MDFDICT	ANS	(creates the formatted dictionary)
MDFDICT	CCT	(creates the formatted dictionary)
MDFDICT	CTW	(creates the formatted dictionary)
MDFWRD50	GLY	(creates formatted dict. for WORD v5.0)
MDFWRD55	GLY	(creates formatted dict. for WORD v5.5)
MDFWRD60	GLY	(creates formatted dict. for WORD v6.0)
MDFENGL	ANS	(creates the finderlist)
MDFENGL	CCT	(creates the finderlist)
MDFENGL	SAV	(creates the finderlist)
MDFENGL1	CCT	(creates the finderlist)
MDFENGL2	CCT	(creates the finderlist)
MDFNATN	ANS	(creates the finderlist)
MDFNATN	CCT	(creates the finderlist)
MDFNATN	SAV	(creates the finderlist)
MDFNATN1	CCT	(creates the finderlist)
MDFNATN2	CCT	(creates the finderlist)
MDFLIST	CCT	(creates the finderlist)
MDFMERG	CCT	(creates the finderlist)
MDFPRT1	CCT	(part of the MDF settings file)
MDFPRT2	CCT	(part of the MDF settings file)
MDFPRT2	SAV	(part of the MDF settings file)
MDFPRT3	CCT	(part of the MDF settings file)
MDFPRT3	SAV	(part of the MDF settings file)
MDFPRT4	CCT	(part of the MDF settings file)
MDFSETT	CCT	(part of the MDF settings file)
MDFSETT	SAV	(part of the MDF settings file)
MDFLANG	CCT	(part of the MDF settings file)
MDFDICT	STY	(MDF stylesheet for dict. and lists)
MDF-FLIP	STY	(changing columns for dict. and lists)
MDF-HP4L	STY	(printing MDF output on HP 4L)
MDF-HP4F	STY	(printing MDF output on HP 4L)
MDF-HPDJ	STY	(printing MDF output on HP Deskjet)
MDF-T321	STY	(printing MDF output on Toshiba 321SL)
MDF-EPLQ	STY	(printing MDF output on Epson LQ series)

G.2 Programs required by MDF

CC	EXE	(Consistent Changes program)
CHOOSE	EXE	(Opening menu program)
CTW	EXE	(Convert to Word program)
FILSPLIT	EXE	(File split program)
SRT	EXE	(Text Analysis Sort program)
TED	COM	(Simple editor)
WORD	EXE	(v5.0 or v5.5, which you must supply)

G.3 Files created by MDF

MDFWORD	GLY	(file for merging split documents)		
SPLIT01	TMP	SPLIT01	DOC	
SPLIT02	TMP	SPLIT02	DOC	
...		...		
DICT	DOC	DICTN01	DOC	(for WINWORD)
DICT	OUT	DICTN02	DOC	(for WINWORD)
DICT	SRT	...		
DICT	TMP	ENGLS01	DOC	(for WINWORD)
ENGL	DOC	ENGLS02	DOC	(for WINWORD)
ENGL	MRG	...		
ENGL	REV	NATNL01	DOC	(for WINWORD)
ENGL	SRT	NATNL02	DOC	(for WINWORD)
ENGL	TMP	...		
NATN	DOC			
NATN	MRG			
NATN	REV			
NATN	SRT			
NATN	TMP			

G.4 Other files included on the release disk

MDFSAMPL	DB	(A sample lexicon, with field markers)
MDFSAMPL	DOC	(Sample formatted as triglot dictionary)
MDFSAMPL	ENG	(English reversed list for sample lexicon)
LXFIELDS	DB	("On-line" helps for field markers—for use in SHOEBOX)
02-START	DOC	(chapter 2 of *Making dictionaries: a guide to lexicography and the Multi-Dictionary Formatter*, providing introductory material, a discussion of all of the field codes, how they are used, and how these standards interact with MDF; earlier DOS version of this chapter)
HP4	STY	(for the START.DOC file)
UPDATE	CCT	(CC table to convert old v0.9x MDF codes to MDF v1.0)
SAGO	PCX	(PCX graphics file for MDFSAMPL.DB)
ANSQ	EXE	(Program useful for tweaking your ANS files)

Appendix H: Macros used in merging process

This section is included for the more technically inclined.

H.1 For WORD v5.0

The following are the macros used in the merging process (WORD v5.0). They are kept in the MDFWRD50.GLY glossary file. This file is copied to MDFWORD.GLY once MDF knows which word processor is being used. It is then renamed to NORMAL.GLY just before WORD is called to merge the split document files. (If there already is a NORMAL.GLY file in the default directory, it is temporarily renamed to MDFXXX.GLY while the documents are being merged. Everything is returned to as before, once the user exits from WORD after perusing the merged document.)

AUTOEXEC MACRO <ctrl a>:

<f6><down 4><end><del><esc>r<space 2>doc<right>.doc<right>n<enter>
<esc>sdoc<down>d<enter><right><shift f6><end><ctrl pgdn><del>
<ctrl pgup><esc>rsplit<right>^«include<space>split<enter>
<esc>r^^p<right>»^^p<enter><esc>fsamdfdict<enter><ctrl r>

REMERGE MACRO <ctrl r>:

«Message Merging split documents, please be patient. »
«SET promptmode = “ignore”»<esc>pmdmdfxxx.doc<enter>
<esc>tlmdfxxx.doc<enter>

H.2 For WORD v5.5

The following are the macros used in the merging process for WORD v5.5. They are kept in the MDFWRD55.GLY glossary file. This file is copied to MDFWORD.GLY once MDF knows which word processor is being used. It is renamed to NORMAL.GLY just before WORD is called to merge the split document files. (If there already is a NORMAL.GLY file in the default directory, it is temporarily renamed to MDFXXX.GLY while the documents are being merged. Everything is returned to as before, once the user exits from WORD, after perusing the merged document.)

AUTOEXEC MACRO <ctrl b>:

<ctrl home><f8><alt e>sSPLIT<tab 3>d<enter><up><end><del>
<alt e>e<space 2>DOC<tab>.DOC<tab 2><space><enter><alt e>sDOC
<tab 3>d<enter><right><ctrl shift f8><end><ctrl end><up><end><del>
<ctrl home><alt e>eSPLIT<tab>^«include<space>SPLIT<tab 2><space>

<enter><alt e>e^^p<tab>»^^p<tab 2><space><enter>
<alt t>amdfdict<enter><ctrl r>

REMERGE MACRO <ctrl r>:

«Message Merging split documents, please be patient. »
<alt f>mnmdfxxx.doc<enter>
<alt f>c<alt f>omdfxxx.doc<enter>

H.3 For WORD v6.0

The following are the macros used in the merging process for WORD v6.0. They are kept in the MDFWRD60.GLY glossary file. This file is copied to MDFWORD.GLY once MDF knows which word processor you are using. It is renamed to NORMAL.GLY just before WORD is called to merge the split document files. (If there already is a NORMAL.GLY file in the default directory, it is temporarily renamed to MDFXXX.GLY while the documents are being merged. Everything is returned to as before, once the user exits from WORD, after perusing the merged document.)

AUTOEXEC MACRO <ctrl b>:

<ctrl home><f8><alt e>sSPLIT<tab 3>d<enter><up><end><del>
<alt e>e<space 2>DOC<tab>.DOC<tab 2><space><enter><alt e>sDOC
<tab 3>d<enter><right><ctrl shift f8><end><ctrl end><up><end><del>
<ctrl home><alt e>eSPLIT<tab>^«include<space>SPLIT<tab 2><space>
<enter><alt e>e^^p<tab>»^^p<tab 2><space><enter>
<alt t>amdfdict<enter><ctrl r>

REMERGE MACRO <ctrl r>:

«Message Merging split documents, please be patient.»«SET echo="off"»
<alt f>mmnmdfxxx.doc<enter>
<alt f>c<alt f>omdfxxx.doc<enter>

Appendix I: Reporting problems or suggesting enhancements

Reports of problems or suggestions for enhancements should be sent to:

JAARS, Inc.
International Computer Services (ICS)
Box 248, JAARS Road
Waxhaw, NC 28173
USA

Telephone: (704) 843–6151
FAX: (704) 843–6200

With any reports of problems please include a printout of the offending entry in its original database format and in its final document form. Please indicate whether the problem or suggestion relates to:

- The MDF program
- The way users interact with the MDF program
- The MDF manual (this *Guide*)
- Your system configuration

With any reports of problems please include your address and a summary of your computer hardware (e.g. Toshiba T1200 with 1Mb memory and 20 Mb hard drive; Toshiba T1950CS with 12Mb of memory and a 200Mb hard drive). Also indicate which version of WORD you are using and which answers you gave to the MDF questions prompted on the screen.

Bibliography

Adelaar, A. K. 1985. Proto-Malayic: the reconstruction of its phonology and parts of its lexicon and morphology. Ph.D. dissertation. Rijksuniversiteit te Leiden. (Published 1992 as *Pacific Linguistics* C–119.)

Apresyan, Yu., Igor Mel'chuk, and A. K. Zholkovsky. 1970. Semantics and lexicography: towards a new type of unilingual dictionary. In Ferenc Kiefer (ed). *Studies in Syntax and Semantics*. Foundations of Language Supplemental Series 10:1–33. Dordrecht: D. Reidel.

———, ———, and ———. 1973. Materials for an explanatory combinatory dictionary of modern Russian. In Ferenc Kiefer (ed). *Trends in Soviet theoretical linguistics*. Foundations of Language Supplemental Series 18:411–438. Dordrecht: D. Reidel.

Bartholomew, Doris A. and Louise C. Schoenhals. 1983. *Bilingual dictionaries for indigenous languages.* Mexico, D.F.: SIL International.

Beekman, John. 1968. Eliciting vocabulary, meaning, and collocation. *Notes on Translation* 29:1–11. Dallas: SIL International. (Reprinted in Alan Healey (ed). 1975. *Language learner's field guide*. Ukarumpa: SIL International. pp. 361–388).

Benson, Morton, Evelyn Benson, and Robert Ilson. 1986. *Lexicographic description of English.* Philadelphia: John Benjamins.

Benson, Morton, Evelyn Benson, and Robert Ilson, compilers. 1986. *The BBI combinatory dictionary of English: a guide to word combinations.* Philadelphia: John Benjamins.

Berlin, Brent, Dennis E. Breedlove, and Peter H. Raven. 1966. Folk taxonomies and biological classification. *Science* 154:273–275.

———, ———, and ———. 1973. General principles of classification and nomenclature in folk biology. *American Anthropologist* 75:214–242.

———, ———, and ———. 1974. *Principles of Tzeltal plant classification: an introduction to the botanical ethnography of a Mayan-speaking people of highland Chiapas.* New York: Academic Press.

Bolton, Rosemary. 1990. A preliminary description of Nuaulu phonology and grammar. M.A. thesis, University of Texas at Arlington.

Bright, William. 1984. The editor's department. *Language* 60:692–693.

Bulmer, Ralph. 1967. Why is the cassowary not a bird? A problem of zoological taxonomy among the Karam of the New Guinea Highlands. *Man* 2:5–25.

———. 1970. Which came first, the chicken or the egg-head? In J. Pouillon and P. Miranda (eds). *Échanges et communications: mélanges offert à Claude Lévi-Strauss a l'occasion de son 60–ième anniversaire.* Paris: Mouton 1970. pp. 1069–1091.

Burchfield, R. W. (ed). 1987. *Studies in lexicography*. Oxford: Clarendon Press.

Carter, Ronald. 1987. *Vocabulary: applied linguistic perspectives.* London: Allen & Unwin.

Casagrande, Joseph B. and Kenneth Hale. 1967. Semantic relationships in Papago folk-definitions. In Dell Hymes and William Bittle (eds). *Studies in southwestern ethnolinguistics.* The Hague: Mouton and Co. pp. 165–193.

Clark, Eve V. and Herbert H. Clark. 1979. When nouns surface as verbs. *Language* 55/4:767–811.

Clynes, Adrian. 1989. Speech styles in Javanese and Balinese. M.A. thesis, Australian National University.

Comrie, Bernard. 1981. *Language universals and linguistic typology.* Oxford: Blackwell.

Comrie, Bernard and Norval Smith. 1977. Lingua descriptive studies: questionnaire. *Lingua* 42:1–72.

Conklin, Harold. 1962. Lexicographical treatment of folk taxonomies. In Fred W. Householder and Sol Saporta (eds). *Problems in lexicography*. pp. 119–141.

Coward, David F. 1990. An introduction to the grammar of Selaru. M.A. thesis, University of Texas at Arlington.

———. 1992–ms. Recommended Maluku lexical database standards. Ambon: SIL International.

Crystal, David. 1985. *A dictionary of linguistics and phonetics.* 2nd edition. Oxford: Basil Blackwell.

Davis, Daniel W. and John S. Wimbish. 1993. *The Linguist's SHOEBOX.* Waxhaw: SIL International.

Dixon, R. M. W. 1979. Ergativity. *Language* 55:59–138.

———. 1982. *Where have all the adjectives gone? and other essays in semantics and syntax.* Amsterdam: Mouton.

———. 1988. *A grammar of Boumaa Fijian.* Chicago: University of Chicago Press.

———. 1991. *A new approach to English grammar, on semantic principles*. Oxford: Clarendon Press.

———. 1994. *Ergativity.* Cambridge Studies in Linguistics 69. Cambridge: University Press.

Durie, Mark. 1985. *A grammar of Achehnese: on the basis of a dialect of north Aceh.* Verhandelingen van het Koninklijk Instituut voor Taal–, Land– en Volkenkunde 112. Cinnaminson, N.J.: Foris Publications.

Ferrell, Raleigh. 1982. *Paiwan Dictionary.* Pacific Linguistics C–73.

Fillmore, Charles J. 1968. Lexical entries for verbs. *Foundations of Language* 4:373–393.

Foley, William A. and Robert D. van Valin, Jr. 1984. *Functional syntax and universal grammar.* Cambridge Studies in Linguistics 38. Cambridge: University Press.

Fox, James J. 1971. Semantic parallelism in Rotinese ritual language. *Bijdragen tot de Taal–, Land– en Volkenkunde* 127:215–255.

———. 1974. 'Our ancestors spoke in pairs': Rotinese views of language, dialect, and code. In Richard Bauman and Joel Scherzer (eds). *Explorations in the ethnography of speaking*. Cambridge: University Press. pp. 65–85.

———. 1975. On binary categories and primary symbols: some Rotinese perspectives. In R. Willis (ed). *The interpretation of symbolism.* ASA Studies 3:99–132. London: Malaby Press.

———. 1977. Roman Jakobson and the comparative study of parallelism. In C. H. van Schooneveld and D. Armstrong (eds). *Roman Jakobson: echoes of his scholarship.* Lisse: Peter de Ridder Press. pp. 59–90.

———. 1982. The Rotinese chotbah as a linguistic performance. *Pacific Linguistics* C–76:311–318.

———. 1988. Introduction. In James J. Fox (ed). *To speak in pairs: essays on the ritual languages of eastern Indonesia.* Cambridge: University Press. pp. 1–28.

Fox, James J. (ed). 1988. *To speak in pairs: essays on the ritual languages of eastern Indonesia.* Cambridge: University Press.

Frake, Charles O. 1962. The ethnographic study of cognitive systems. In *Anthropology and human behavior.* Washington D.C.: Anthropological Society of Washington. pp. 28–41.

Franklin, Karl. 1992. Lexicography considerations for Tok Pisin. Paper presented at the Congress of the Linguistic Society of Papua New Guinea, September 1992. Madang.

Givón, Talmy. 1984. *Syntax: a functional-typological introduction,* Vol. 1. Amsterdam: John Benjamins.

———. 1990. *Syntax: a functional-typological introduction,* Vol. 2. Amsterdam: John Benjamins.

Gleason, H.A. Jr. 1962. The relation of lexicon and grammar. In Householder and Saporta (eds). *Problems in lexicography*. pp. 85–102.

Grace, George. 1981. *An essay on language.* Columbia, S.C.: Hornbeam Press.

———. 1987. *The linguistic construction of reality*. Sydney: Croon Helm.

Grimes, Barbara Dix. 1991. The development and use of Ambonese Malay. *Pacific Linguistics* A–81:83–123.

Grimes, Barbara F. (ed). 1992. *Ethnologue: languages of the world.* 12th edition. Dallas: SIL International.

Grimes, Charles E. 1987. Mapping a culture through networks of meaning. *Notes on Linguistics* 39:25–46.

———. 1991. The Buru language of eastern Indonesia. Ph.D. dissertation. Canberra: Australian National University.

———. 1992. Refining parts of speech in the lexicon. Paper presented at 1992 Asia International Lexicography Conference, October 1992. Manila.

———. 1994. Mapping semantic relationships in the lexicon using lexical functions. *Notes on Linguistics* 65:5–25.

Grimes, Charles E. and Kenneth Maryott. 1994. Named speech registers in Austronesian languages. In Tom Dutton and Darrell T. Tryon (eds)., *Language contact and change in the Austronesian world.* Trends in Linguistics Studies and Monographs 77. Berlin: Mouton de Gruyter. pp. 275–319.

Grimes, Joseph E. 1980a. Huichol life form classification: I—Animals. *Anthropological Linguistics* 22:187–200.

———. 1980b. Huichol life form classification: II—Plants. *Anthropological Linguistics* 22:264–274.

———. 1989. Information dependencies in lexical subentries. In. M. W. Evens (ed). *Relational models of the lexicon: representing knowledge in semantic networks.* Cambridge: University Press. pp. 167–182.

———. 1990. Inverse lexical functions. In J. Steele (ed). *Meaning–Text Theory: linguistics, lexicography, and implications.* University of Ottawa Press, Ottawa. pp. 350–364.

———. 1992. Lexical functions across languages. In *Proceedings of the International Workshop on The Meaning–Text Theory*, 27 July – 3 August 1992. Darmstadt, Germany. pp. 123–131.

———. 1987–ms. A field guide to words: relations and linkages in the lexicon. Dallas: SIL International.

Grimes, Joseph E. and Barbara F. Grimes. 1993. *Ethnologue language family index.* Dallas: SIL International.

Grimes, José, and others. 1981. *El Huichol: apuntes sobre el léxico*. Department of Modern Languages and Linguistics, Cornell University, Ithaca, NY. [Out of print, reissued as ERIC document ED 210 901].

Haiman, John. 1980. Dictionaries and encyclopaedias. *Lingua* 50:329–357.

Halliday, M. A. K. 1961. Categories of the theory of grammar. *Word* 17:241–292.

Hartmann, Reinhard R. K. (ed). 1983. *Lexicography: principles and practice.* London: Academic Press.

———. 1986. *The history of lexicography.* Philadelphia: John Benjamins.

Hashimoto, Mantaro J. 1977. *The Newari language: a classified lexicon of its Bhadgaon dialect.* Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa.

Horne, Elinore Clark. 1974. *Javanese-English dictionary*. New Haven: Yale University Press.

Householder, F.W. and Sol Saporta (eds). 1962. *Problems in lexicography*. Bloomington: Indiana University Research Center in Anthropology, Folklore and Linguistics.

Hughes, Jock, 1991–ms. Dobel, a language of the Aru Islands. Ambon: Pattimura University and SIL International.

Ilson, Robert (ed). 1987. *A spectrum of lexicography*. Philadelphia: John Benjamins.

Jacobson, Marc. R. 1986. *Philippine dictionaries on computer.* Manila: SIL International.

Lakoff, George. 1987. *Women, fire, and dangerous things: what categories reveal about the mind.* Chicago: University of Chicago Press.

Lakoff, George and Mark Johnson. 1980. *Metaphors we live by.* Chicago: University of Chicago Press.

Lakoff, George and Mark Turner. 1989. *More than cool reason – a field guide to Poetic Metaphor.* Chicago: University of Chicago Press.

Landau, Sidney I. 1989. *Dictionaries: the art and craft of lexicography.* Cambridge: University Press.

Langacker, Ronald, W. (ed). 1977–1984. *Studies in Uto-Aztecan grammar,* Vols. 1–4. Dallas: SIL International and University of Texas at Arlington.

Leed, Richard L. and Alexander D. Nakhimovsky. 1979. Lexical functions and language learning. *Slavic and East European Journal* 23(1):104–113. [Revised in J. Steele (ed). 1990. *Meaning–Text Theory: linguistics, lexicography, and implications.* Ottawa: University of Ottawa Press. pp. 365–375].

Lehmann, Christian. 1982. Directions for interlinear morphemic translations. *Folia Linguistica* 16:199–224.

Louw, Johannes P. and Eugene A. Nida (eds). 1988. *Greek–English lexicon of the New Testament based on semantic domains*. New York: United Bible Societies.

McKeon, Richard (ed). 1941. *The basic works of Aristotle*. New York: Random House.

Mel'chuk, Igor, 1973. Towards a linguistic "meaning–text" model. In Ferenc Kiefer (ed). *Trends in Soviet theoretical linguistics*. Foundations of Language Supplemental Series 18:35–57. Dordrecht: D. Reidel.

———. 1982. Lexical functions in lexicographic description. In *Proceedings of the Eighth Annual Meeting of the Berkeley Linguistics Society.* Berkeley: Department of Slavic Languages and Literatures, University of California. pp. 427–444.

———. 1989. *Explanatory Combinatorial Dictionary and Learner's Dictionaries*. SEAMEO Regional Language Centre, Occasional Papers No. 45. Singapore: RELC

Mel'chuk, Igor and Nikolaj V. Pertsov. 1986. *Surface syntax of English: a formal model within the meaning–text framework.* Philadelphia: John Benjamins.

Mel'chuk, Igor and Alain Polguère. 1987. A formal lexicon in meaning–text theory (or how to do lexica with words). *Computational Linguistics* 13(3/4):261–275.

Mel'chuk, Igor and A.K. Zholkovsky. 1970. Towards a functioning meaning–text model of language. *Linguistics* 57:10–47.

——— and ———. 1984. *Explanatory combinatorial dictionary of modern Russian.* Vienna: Wiener Slawistischer Almanach.

——— and ———. 1988. The Explanatory Combinatorial Dictionary. In. M. W. Evens (ed). *Relational models of the lexicon: representing knowledge in semantic networks.* Cambridge: University Press. pp. 41–74.

Moore, Bruce R. *Doublets in the New Testament*. Dallas: SIL International.

Mosel, Ulrike. 1991. Markedness theory and the distinction of major word classes in Samoan. Seminar presented at the Australian National University. Canberra.

Murdock, George, and others. 1982. *Outline of cultural materials.* 5th revision. New Haven, Connecticut: Human Relations Area Files, Inc.

Newell, Leonard E. 1986. Lexicography notes. Typescript. Manila: SIL International.

———. 1993. *Batad Ifugao dictionary: with ethnographic notes.* Manila: Linguistic Society of the Philippines.

Nida, Eugene. 1949. *Morphology*. Ann Arbor: University of Michigan Press.

———. 1958. Analysis of meaning and dictionary making. *International Journal of American Linguistics* 24:279–292.

Nothofer, Bernd. 1982. Central Javanese dialects. *Pacific Linguistics* C–76:287–309.

Pawley, Andrew K. 1973. Some problems in Proto-Oceanic grammar. *Oceanic Linguistics* 12(1/2):103–188.

———. 1986. Lexicalization. In Deborah Tannen and James E. Alatis (eds). *Languages and Linguistics: the interdependence of theory, data, and application.* Georgetown University Round Table on Languages and Linguistics, 1985. Washington, D.C: Georgetown University Press. pp. 98–120.

———. 1993. Lecture notes: dictionaries and dictionary making. Canberra: Department of Linguistics, The Australian National University.

Poedjosoedarmo, Soepomo. 1968. Javanese speech levels. *Indonesia* 6:54–81.

Robinson, Dow F. 1969. *Manual for bilingual dictionaries*. Santa Ana, California: SIL International.

Ross, Malcolm D. in press. Reconstructing Proto Austronesian verbal morphology: evidence from Taiwan. Paper presented at International Symposium on Austronesian Studies relating to Taiwan. December 1992.

Schachter, Paul. 1976. The subject in Philippine languages: topic, actor, actor-topic or none of the above? In Charles Li (ed). *Subject and Topic*. New York: Academic Press. pp. 491–518.

———. 1977. Reference-related and role-related properties of subjects. In Cole and Sadock (eds). *Syntax and semantics 8: grammatical relations.* New York: Academic Press. pp. 279–306.

———. 1985. Part-of-speech systems. In Timothy Shopen (ed). *Language typology and syntactic description I: clause structure.* Cambridge: University Press. pp. 3–61.

Starosta, Stanley, Andrew K. Pawley, and Lawrence A. Reid. 1982. The evolution of focus in Austronesian. *Pacific Linguistics* C–75:145–170.

Steele, James (ed). 1990. *Meaning–Text theory: linguistics, lexicography, and implications.* Ottawa: University of Ottawa Press.

Simons, Gary F. 1979. *Language variation and limits to communication*. Ithaca, N.Y.: Department of Modern Languages and Linguistics, Cornell University.

Simons, Gary F. and Larry Versaw. 1987. *How to use IT: a guide to interlinear text processing.* Dallas: SIL International.

Svenson, B. 1992. *Practical lexicography: principles and methods of dictionary making.* Oxford: Oxford University Press.

Taumoefolau, Melenaite. 1991. Verbal senses of concrete nouns in Tongan. Paper presented at the Sixth International Conference on Austronesian Linguistics, May 1991. Honolulu, Hawai'i.

Therik, Tom and Charles E. Grimes. 1992–ms. Baria Ulu: a Tetun text. Canberra: Australian National University.

Tomaszczyk, Jerzy, and Barbara Lewandowska-Tomaszczyk (eds). 1990. *Meaning and lexicography.* Philadelphia: John Benjamins.

Vonen, Arnfinn M. 1991. Hunting for nouns and verbs in Samoan. Seminar presented at the Australian National University, 22 November 1991. Canberra.

———. 1992. Nominalisations in Tokelau. Seminar presented at the Australian National University, 15 May 1992. Canberra.

Weinrich, Uriel. 1962. Lexicographic definitions in descriptive semantics. In Householder and Saporta (eds). *Problems in lexicography.* pp. 25–44.

Wierzbicka, Anna. 1980. *Lingua mentalis: the semantics of natural language.* New York: Academic Press.

———. 1985. *Lexicography and conceptual analysis*. Ann Arbor: Karoma Publishers.

———. 1986. What's in a noun? (Or: How do nouns differ in meaning from adjectives?) *Studies in Language* 10(2):353–389.

———. 1988. *The semantics of grammar.* Studies in Language Companion Series 18. Amsterdam: John Benjamins.

———. 1991. *Cross-cultural pragmatics: the semantics of human interaction*. Trends in Linguistics Studies and Monographs 53. Berlin: Mouton de Gruyter.

———. 1992. *Semantics, culture, and cognition: universal human concepts in culture-specific configurations.* Oxford: Oxford University Press.

———. to appear-a. Adjectives vs. verbs: the iconicity of part of speech membership. In: M. Landsberg (ed). Proceedings of a symposium on iconicity. Zagreb.

———. to appear-b. Back to definitions: cognition, semantics, and lexicography. In *Lexicographica* 8.

Wimbish, John S. 1989. *Shoebox: a data management program for the field linguist*. Waxhaw: SIL International.

Wolff, John, and Soepomo Poedjosoedarmo. 1982. *Communicative codes in Central Java.* Ithaca, N.Y.: Southeast Asia Program, Cornell University.

Wurm, Stephen A. and B. Wilson. *English finderlist of reconstructions in Austronesian languages (post-Brandstetter).* Pacific Linguistics C–33.

Zgusta, Ladislav. 1971. *Manual of lexicography*. The Hague: Mouton.

Zgusta, Ladislav (ed). 1980. *Theory and method in lexicography: Western and non-western perspectives.* Columbia, S.C: Hornbeam Press.

———. 1988. *Lexicography today: an annotated bibliography of the theory of lexicography.* Max Niemeyer Verlag: Tubingen.

Index

—A—

abbreviations 15, 24, 37, 43, 124, 172, 175, 180, 195
abstract terms 68
academic audience 68, 140, 165, 180
acknowledgments 181
active intransitive 167
active transitive 167
active verbs 166
activities 130
activities and events 151
Actor 152, 166
Actor noun 127
actors 21
Adelaar 165
adjectives 15, 160, 170, 171, 192
adpositions 161, 162
Adult 130
affixes 51, 103, 159, 163
agent 152
all-purpose fields 21
alphabetizing 67, 89, 93, 104
alternate pronunciations 23
ambiguity 112
ambitransitive 169
ambivalent category 163
anaphoric pointers 107
animals 68, 141, 144
Ant 133, 134
anthropologist 137
Anti 133
antonym 122
antonyms 21, 22, 101, 102, 132, 133, 202
applying a style in WORD 65
Apresyan 121, 123
archaic words 40
archiving dying languages 67
Aristotle 137
artifacts 73
associated activities 143, 145
asterisk 17, 42
attributive 170
audience 68, 77, 104, 157, 178, 187
AUTOEXEC.BAT 2, 55
automated reverse indexing 13
automatic pagination 54
autosave 54
avifauna 19

—B—

backslash codes 9
back-up 5
Bartholomew and Schoenhals 19, 105, 106, 107, 115, 157
basic field markers 16
basic set of fields 76
basic strategies 67
beginning of a dictionary project 177
Benefactee 128
Berlin, Breedlove and Raven 142
bibliographical references 27, 93, 205
bilingual 41, 71
bilingual dictionaries 15, 16, 60, 67, 70, 71, 105, 114, 117, 148, 158
Birds 146
body part terms 67, 68, 69, 96, 115, 148, 191
Bolton 168
borrowed words 24, 113, 153, 178, 204
botanists 73, 137, 141
botany 19
both a noun and a verb 161
bound morphemes 13, 42, 81, 95, 165
bound roots 14, 86, 93, 95, 104, 164
Bright 173
Bulmer 142
bundles 21

—C—

candidates for headwords 99
Cap 133
carrying verbs 74, 115, 116
Casagrande and Hale 142
categories of information in a lexical entry 92
categorization 157
category labels 158
Caus 131
Causal 131
causative 153

CAUTION ..... 1, 4, 13, 14, 15, 17, 20, 27, 53, ... 73, 74, 107, 110, 111, 112, 113, 114, ... 140, 158, 161, 164, 192, 200
CC table 56, 58, 60, 200
Cess 132
Cessative 132
CHANGE SETTINGS 9, 56, 199
change-of-states 152
changes in field markers 200
character formatting codes 49, 50, 199, 200, 206
character styles 49, 50, 51, 58, 64, 206
chart 25
check for consistency 71
checking senses 73
chevrons 52
Child 126, 130
choosing example sentences 57, 105, 106
choosing headwords 99
circular 40
citation form ..... 13, 14, 58, 78, 86, 93, 96, ... 104, 105, 171
classifier system 139
clichés 102
cliticized forms 23
cluster of properties 158
Clynes 154
collective 133
collective knowledge 142
combination of keys 3
combinatory possibilities 159
command line 1, 54
comments 20, 23, 24, 28, 154
comments related to any field 28
commercial dictionaries 9, 60
community involvement 177
community leaders 69, 181
Comp 132
compacted 57
comparative and historical linguistics 154
comparative linguists 118, 153
compiler 68
compiler-centric 157
complement 132
complementary 134
complementary distribution 45, 159, 162
completing the dictionary 177
composite functions 193
Compound 130
compounds 22, 67, 73, 99, 100, 102
compromise 84
Computer Assisted Related Language Adaptation [CARLA] programs 118
computer software manual 3
computerized graphic 27
computerized lexical database 70, 74
Comrie 173
conceptual correspondence 140
concordance 109
confer 22
conjunctions 160, 161, 162
Conklin 142
connotative meaning 39
Conseq 128
consequence 128
consistency in labeling 8, 15, 83, 175
Consistent Changes [CC] program 200
content words 164
contexts 36
contextual meaning 115
contrastive patterns 158
Conv 132
conventionalized knowledge 80, 101
converse 132
Convert-to-Word [CTW] program 58
co-occurrence restrictions 105
core arguments 171
corpus of natural texts 73
corrupted file 5
Counterpart 132, 134
counting headwords 67
Coward 167
Cpart 132, 134
cross-reference ..... 4, 14, 21, 22, 23, 49, 64, 67, ... 79, 82, 83, 94, 100, 119, 125, 126, 139, ... 180, 202
Crystal 39
cultural items 150
cultural-linguistic units 84, 99, 101
customize 134
customize the output 13
customized output 13
customized primary sort sequences 93
cutting verbs 74, 116, 126

—D—

data management 67
data notebooks 19

database format ... 9
database structure ... 7, 9, 89
database template ... 75, 76, 122
data-gathering methods ... 72
Date ... 29
decayed state ... 131
default audience ... 68
default configuration ... 54
default sort order ... 58
definitions ... 16, 17, 18, 19, 36, 38, 39, 40, 41, 45, 70, 71, 105, 114, 137, 138, 150, 164
Degrad ... 131
deictics ... 38, 159
denotative meaning ... 39, 45, 155
department of education ... 84
description ... 16, 18
deteriorated state ... 131
determining parts of speech ... 158
deverbal noun ... 128
diacritics ... 179
dialect information ... 117, 120
dialect map ... 179
dialect names ... 20, 22, 24, 119, 120
dialect variants ... 23, 118, 119, 179
dialectal synonyms ... 119, 124, 155, 179
dialectal variants ... 23
dialects ... 119, 171, 179
dictionaries ... 60
dictionary ... 4, 5, 7, 9, 14, 16, 28, 42, 43, 66, 67, 68, 69, 77, 84, 118, 161, 177, 180
dictionary of a related language ... 73
dictionary users ... 39
dictionary-making ... 7
different audiences ... 13
different classes of notes ... 28
different distributional networks ... 118
different meanings ... 118
different purposes ... 36
different senses ... 107, 109, 110, 111, 112, 114
differentiae ... 40, 137
diglot ... 15, 34, 64, 71, 199
digraphs ... 6, 7, 58, 89, 93, 94, 95
diminished degree ... 131
directionals ... 38
disadvantages ... 89
discarded ... 17
discarding fields ... 57
discourse particle ... 162
disease ... 68
distinguishing usage restrictions ... 23
distribution ... 117, 120, 159, 162, 164
division breaks ... 95
Dixon ... 166, 169, 170, 191
dot on the screen ... 58
dot-matrix printer ... 63
double quotes ... 52
dual ... 25
duplicate glosses ... 9
Durie ... 167

—E—

edible plants ... 68
editorial changes ... 89
em-dash ... 45
emic ... 140
emic units ... 100
emic unity ... 36
emic vernacular categories ... 27
emotion words ... 74
emotions ... 74
empty \lf fields ... 21
encyclopedic fields ... 18, 135, 205
encyclopedic information ... 20, 39, 137, 200
English finderlist ... 53
enhancements and changes to MDF ... 199
entry ... 16, 17, 21, 28, 42, 180, 203
Equip ... 133
ergative ... 166
ethnobotanists ... 137
ethnographic information ... 20
ethnographic notes ... 201
ethnographic sketch ... 180
ethnolinguistic pride ... 69
etic ... 140
etic checklist ... 27
etymology ... 24, 70, 153, 163, 178, 204
events ... 152
example sentences ... 16, 19, 67, 70, 71, 105, 199
examples extracted from texts ... 19
examples from dictionaries ... 4
excessive duplication ... 43
exclude certain fields ... 56, 57
exclude entries ... 28
exclude from the reversed finderlists ... 42
exclude part of speech ... 61
excluding example sentences ... 57

excluding your notes 57
exclusive 25
expanded entries 74
expanded glosses 38
experiencer 152
explanation 18
extended sense 149
extracting topical subsets 26, 89, 177

—F—

false polysemy 114
fast searches 7
fauna 19, 73, 137, 140, 141, 142
Feel 132
Female 126
Ferrell 164
field codes 1, 13
field markers 183
field researchers 60
FIESTA 109
figurative sense 102
files and programs 207
files created 208
filter 122
Filters 8
Fin 132
Final phase 132
final punctuation 15
financial resources 70
finderlists 5, 17, 41, 43, 56, 60, 61, 64, 67
finding words 72
first gloss 17
fish 19, 67, 68, 147
fish names 115
fixed format 25
floppy drive 2, 55
flora 19, 73, 137, 140, 141, 142
Foley & Van Valin 166
Foley and Van Valin 116, 173
folk etymologies 110
folk taxonomies 25, 27, 125, 126, 138
footers 63
form 158
form class 157
formalism 39
Format dictionary 53, 56, 57
formatted dictionary 10, 54
formatted output 10
formatting 9, 67
Fox 103, 156
Frake 142
free disk space 55
free translation 19
free-form fields 49, 51
from the beginning 4
fully edited 29
function 158
functors 41, 51, 155, 161, 162, 172, 173
fv: 10, 49, 50, 51, 206

—G—

Gen 125
gender 25
general audience 69
general note 28
generic 25, 27, 68, 80, 125, 126, 139, 151
generic-specific 21, 112, 125, 126
genus 40, 137
Givón 157, 169, 173
gloss 16, 17, 18, 36, 67, 90
gloss fields 16, 36, 37, 38, 41, 187
glossary 67, 69, 209, 210
glossary files 2, 54
glosses 70
glossing strategies 36
goal 128
government authorities 66
gradation 132
grammatical introduction 153, 162
grammatical paradigm 25
grammatical particles 37
grammatical restrictions 21, 105
graphics format type 28
Grimes and Maryott 155
Grimes, B.D. 113
Grimes, B.F. 179
Grimes, C. 96, 116, 121, 123, 155, 159, 162, 167, 168, 170, 201
Grimes, J. 122, 123, 124, 142, 193
Grimes, J. and B.F. Grimes 179
Group 133
group exploration 142

—H—

Halliday 164

hanging indents 30
hard copy printout 5
Hashimoto 27
Head 133
headers 63
headword 13, 14, 16, 18, 19, 22, 40, 67, 73, 79, 89, 92, 96, 99, 101, 105, 125, 150, 205
helps file 13
hierarchical structure of an entry 45
high frequency words 40
historical and comparative linguistics 113
historical reconstructions 154
historically related 112
homograph 14
homonym number 58
homonym numbers 23, 57, 58
homonyms 14, 22, 45, 58, 61, 83, 93, 94, 109, 110, 111, 113, 162, 163, 180
homonymy 107, 109, 115
homonymy, partial 169
homophone 9, 14
Horne 40, 154, 155
housekeeping field 19, 24
housekeeping fields 28
housekeeping information ... 8, 29, 67, 75, 89, 93
houses 74, 150
HRAF 27
Hughes 167
Human Relations Area Files 27
hyperonym 125
hyponym 126

—I—

identify polymorphemic words 73
idioms 19, 67, 101, 102, 103, 134, 148, 153
ignore your notes fields 199
ignored for reversal 17
illustrative sentences 19, 105, 106, 107
immature phase 141
imperfective 167
Incep 131
inceptive 131
inchoative 131
inclusive 25
incomplete inflections 23
Incr 130
indefinable 40
Indefinite terms 174
index 67
index of semantics 27, 115
indexed by the root 41
Indiv 133
infinitives 41, 96
infix 104
inflected for person and number 78
inflected forms 96
information about the headword 92
inherent meaning 115
inherited vocabulary 113, 153
initial phase 131
inkjet printers 63
insects 68, 147
installing MDF 1
institutionalized status 102
instrument 21, 128, 152
interaction with language assistants 123
interlinearize 44, 90
interlinearizing 8, 17, 18, 20, 36, 37, 41, 43, 73, 75, 81, 84, 90, 103
intermediate taxa 138, 140, 141
internal fields 9
intradirective verbs 167, 168
intransitive 160
introduction to the dictionary 6, 14, 25, 78, 118, 178
irregular paradigms 25
isolating language 99

—J—

Javanese 40
joining underline 17
Jump feature 7, 144
jumping to nonadjacent entries 89
jungle plants 68

—K—

key field 13, 58, 104, 201
keyboard conventions 3
keyboard setup 2, 54
kin terms 38, 67, 68, 115, 148, 149, 177
kinship 89, 150, 180, 198
knowledge bank 20

—L—

Lakoff 142
Landau 9, 77, 115
Langacker 173
language code 50
language community 66
language learners 118
language of parallelism 103
large print job 63
Lead 133
learn the language and culture 72
Lehmann 173
lemma 13
lexeme 13, 19, 38, 67, 100, 101, 106, 161, 205
lexeme-based 78, 79, 82, 83, 84
lexeme-oriented 77, 78
lexical associations 121
lexical citation form 14, 61
lexical database 5, 9, 13, 54, 60, 67, 71, 73, 75, 84, 103, 118, 142
lexical entry 15, 43, 61, 73, 92
lexical functions 16, 20, 21, 106, 110, 121, 123, 134, 135, 193
lexical networks 74, 101, 110, 121, 141
lexical relations 121, 201
lexical roots 164
lexical sets of similar words 72
lexical universals 39, 40
lexicalized 101
lexicalized circumlocutions 125
lexicalized compounds 130
lexicographers 15
lexicography 3, 7
lexicon 67
LEXICON.DB 6, 54
life forms 138, 140, 141
limitations 54
lingua franca 18, 72, 113, 153
linguistic analysis 75
Liqu 132
literally 19
literature 27
loan sources 198
loan synonym 124, 179
loans 24, 113, 124, 153
local audience 69, 104, 105, 165
local audiences 77
local community 20, 68, 69, 71, 83, 140, 165, 177
local government 69
local population 69
location 127, 152
long headwords 62
look up this word 36
loose definitions 38
Louw and Nida 27
LXFIELDS.DB 4

—M—

MACROS 52, 76, 122, 204, 209
Magn 130
main entry 19, 22, 23
major word classes 40
Male 126
Maluku Dictionary Formatter 200
Manif 132
mapping lexical networks 21
margins 63
Mat 129
Material 129, 150
material culture 73
material world 150
mature phase 141
Max 130
maximalist 137
McKeon 137
MDF fields 13
MDF files 1
MDF output 29
MDFDICT.ANS 94
MDF-prompted options 56
MDFSAMPL.DB 4, 53
meaning 18, 36, 38, 39, 67, 114, 115, 121
meaning-centric 77
meaning-oriented 79
medicines 68
Mel'chuk 121, 123, 124
menu options 9, 53, 56
metaphors 151
metathesis 24, 154
Min 131
minimal entries 67, 74
minimalist 137
minor entries 23, 41, 42
minor sense 16, 165

minor variant 22, 23
mismatch of terms 72
mixed audience 69
modify the default settings 56
modifying the printout 64
monolingual dictionary 16, 40, 67, 70, 71
monomorphemic 84, 99
monospace font 3, 14
Moore 156
more than one bundle 203
more than one \ps 15
more than one sense 16
more than one version of WORD 55
morpheme breaks 174
morpheme representation 22
morpheme-by-morpheme 205
morpheme-level 17, 18, 38, 81
morphemic arrangement 77
morphological causative 153
morphological variants 23
morphologically complex national language ... 41
morphologically defined subclasses 168
morphology 22
morphophonemic processes 22, 179
morphosyntactic network 157, 159, 163, 168, 171
Mult 133
Multi-Dictionary Formatter program 53
multilingual bundles of field markers 90
multilingual databases 90
multiple bundles 21
multiple criteria 141
multiple examples 19
multiple glosses 17, 37, 105
multiple language information 8, 90
multiple parts of speech 45
multiple senses 14, 15, 16, 45, 67, 108
multiple word glosses 17
Murdock 27

—N—

Nact 127
naive user 15
national audience 15
national government 69
national language 18, 20, 34, 49, 50, 54, 64, 69, 72, 90, 113, 120, 153, 154, 158, 187
national language dictionaries 15
national language institute 83
native nomenclature 140
native speakers 16, 41, 72, 74, 107, 109, 121, 125, 142
native taxonomy 126, 151
natural environment 151
natural semantic metalanguage 39
natural text 107, 109
Nben 128
Ndev 128
near synonyms 110, 126
needing editing 29
networks of meaning 121
Newell 115, 180
Ngoal 128
Nida 157
Ninst 128
Nloc 127
no content in a field 76
nomenclature 138
nominal argument 127
non-active verbs 167
non-adjacent entries 7
non-animate 25
non-core arguments 159
non-human 25
non-native speaker 16, 107
non-printing characters 95
non-restrictive 99
not recognized by MDF 29
NOTE 4, 18, 42, 53, 57, 62
Note fields 28
Nothofer 155
noun class 25
nouns or verbs? 162
Nug 127

—O—

odd-even running footers 58
On-line helps 4
Only 21
order of fields 4, 13, 187
Organization 133
original lexical database 53
orthographic conventions 52, 179
output file 58
over-differentiated 141

—P—

paradigms 23, 25, 171, 175
parallelisms 134, 156
paraphrase test 110
ParD 134
ParS 134
parse words 105
parsing 73
Part 129
part of speech 15, 40, 45, 62, 67, 75, 109, 115
partial homonymy 108, 163
particles 51
parts of speech ... 16, 37, 109, 157, 159, 175, 195
part-whole 21, 112, 141
path 1, 2, 54, 55
patient 152
Pawley 71, 72, 74, 100, 101, 103, 114, 115, 137, 150, 168
PCX 27
perfective 167
periphrastic causative 153
Perm 131
Phase 130, 141
phonetic 204
phonetic fonts 14
phonetic form 14
phonotactically similar 134
photograph 27
phrasal lexemes 13, 73, 100, 102
phrasal units 67
phrases 49
physical characteristics 141, 142, 146
picture 28
picture books 73
picture in entry 27
plain space 17
plant names 115
plants 67, 74, 141, 142, 177
plural 25
Plus 130
Poedjosoedarmo 154
poetic text 134
political considerations 83
polymorphemic 79, 81, 82, 83, 84, 179, 205
polymorphemic forms 14, 47, 81
polysemy 107, 109, 114, 115, 148
polysynthetic language 99
portmanteau morphemes 174
post-editing 6, 83, 94, 95
postpositions 162
practical orthography 14
pragmatic connotations 20
pragmatically motivated variants 169
precategorials 96, 164, 165
preceding hyphens 95
predicative 170
prefixes 104, 158
prefixing languages 77, 104
preliminary volume 68
Prep 130
preparatory activity 130, 152
prepositional verbs 159
prepositions 159, 162
prestige 69, 71
presupposed information 107
primary audience 68
principles 40, 73, 99, 109, 158
print tables 207
printing 63
printing the dictionary 17
processes 152
proclitics 158
programs required 208
pronouns 38, 173
pronunciation 14
propositions 140
prose explanations 38
proto forms 24, 204
Prox 131
publication 178
publishing costs 137
punctuation 10, 50, 52
purpose 77, 90, 178

—Q—

qualities 152
quality control 8
quasi-reflexive verbs 167, 168
Quit 62

—R—

range of functions 158
range of meaning 18, 67, 70, 109, 115, 150
Range sets 8, 15, 175
raw SHOEBOX form 29

recommendation 70, 78
reconstructed forms 24
record marker 13
redundant information 41
reduplication 23, 25, 153, 174
reference 19
referential meaning 39
referential prominence 169
refine entries 109
reflexive 168
region 20
regional creoles 18
regional language 18, 20, 34, 50, 90, 120, 187
register 20
register synonym 125, 179
related languages 66
related lexical entries 82
relater 162
release disk 53, 199
reliability of the information 93
reptiles 68
requirements 54
requirements and limitations 2
Res 128
researcher's national language 92
Reset 57
Reset option 199
restores the settings file 57
restrictions 21, 120, 179
Result 128
resulting state 128, 152
Rev 133
reversal 17, 18, 19, 36, 37, 41, 67
reversal fields 37, 90
reverse the glosses 60
reversed finderlist 5, 9, 10, 76
reversed finderlists 14, 16, 17, 36, 41, 180
reversed index 60
reversing the dictionary 89
ritual language 69, 103
ritual speech 154, 156
root morphemes 14
root-based 47, 78, 79, 83, 84
root-based database 81
root-oriented 77, 78
Ross 164
running MDF 1

—S—

safekeeping 5
same meaning 118
sample database 4, 54
sample file 53
scale 132, 133
Schachter 157, 159, 170
scholarly audience 70, 104
scientific name 19, 50, 73
scientific nomenclature 73, 140, 141
scientific taxonomy 140
scope 162
screen prompts 9
search and retrieval 122
secondary sort character 95
secondary sort order 93
semantic arrangement 77
semantic categories 26, 115
semantic domain 26, 27, 37, 68, 73, 74, 115, 175, 177, 191
semantic primitives 39, 40
semantic shift 113, 117, 154
semantically bleached senses 99
semantically complex things 40
semantically related entries 27
sensation 132
sense 9, 17, 19, 21, 28, 180, 203
sense discrimination 105
sense number 16, 45, 61
sense numbers 45
sentence number 19
separate dictionaries 69
separate publications 71
separate volumes 177
Seq 130
sequence of key strokes 3
Serial 129
serial verbs 159
sets of similar words 73
several researchers 28
shared meaning 109, 110, 111
shared semantic thread 109
SHOEBOX 9, 13, 26, 53, 56, 57, 58, 60, 73, 75, 76, 89, 93, 115, 122, 142, 144, 175
SHOEBOX's Jump feature 82
SHOEBOX datestamp 29
SHOEBOX Filters 27, 90, 177
SHOEBOX interlinear function 17

Sim ... 126
similar ... 126
Simons ... 117
Simons and Versaw ... 172
simple morphemes ... 100
simple reversals ... 67
Sing ... 133
singular ... 25
Sit ... 130
situations ... 130
sketch in a notebook ... 27
slide ... 27
small caps ... 3
social usage ... 120
sociolinguistics ... 120, 155
Son ... 132
sort ... 14
sort order ... 58
sort sequences ... 93
SORT.EXE ... 55
sorting ... 89, 93
Sound ... 132
source language ... 24
source of data ... 28
space-semicolon-space ... 37
spacing integrity ... 17
Spec ... 126
special characters ... 52
special classes of entries ... 137
special registers ... 154
specialized dictionaries ... 67
species ... 40, 126, 137
specifics ... 126, 139, 151
speech register name ... 22
speech registers ... 125, 154, 155
speech-act verbs ... 74, 115
speed in interlinearizing ... 17
spelling variants ... 120
split document files ... 59
split intransitive ... 166
split-S ... 160, 166
sponsoring agencies ... 177
SRT.EXE ... 58
standard field codes ... 53
standard format markers ... 9
Starosta, Pawley, and Reid ... 164
Start ... 131
starter list ... 191, 193, 195
State ... 132
states and processes ... 152
stative-active ... 166, 167
status for editing ... 28
stimulating community interest ... 68
Stop ... 132
strategies of abbreviations ... 172
structural hierarchy of an entry ... 15
structural variants ... 120
structure-centric ... 77
structuring entries ... 49
structuring information ... 99
stylesheet ... 5, 9, 63, 64
Subadult ... 126
subcategories ... 160
subentries ... 16, 47, 82, 100, 108, 165, 180
subentry ... 14, 16, 17, 19, 21, 28, 42, 61, 83, 203, 205
subscripts ... 14, 94
subsenses ... 46
substantive ... 127
suffixes ... 95, 104
suffixing languages ... 104
Super ... 130
superlative degree ... 130
surface form ... 14, 78, 96
switch-reference ... 161
symbol [§] ... 4
Sympt ... 132
Syn ... 124
SynD ... 124
SynL ... 124
synonyms ... 21, 22, 67, 101, 102, 110, 124, 202
SynR ... 125
SynT ... 125
syntactic behavior ... 159
syntactic classes ... 151
syntactic slots ... 161

—T—

tables in an entry ... 25
taboo synonym ... 125, 179
taboos ... 145
taxonomy ... 19, 27
team of compilers ... 28
technical definitions ... 18, 70
technical jargon ... 40, 68
TED.COM ... 56
Template ... 8, 76

terminal taxa 138, 139, 140
terminological correspondence 141
terminological system 141
terminology 67
test file 4
test MDF 2, 95
Text Analysis [TA] program 58
text corpus 8
text name 19
text only 53, 54, 95
text-based lexicography 8, 72
Therik and Grimes 171
thesaurus 27, 68, 115
TIP ..... 4, 16, 21, 41, 42, 50, 51, 56, ... 75, 90, 93, 106
transitive 160
translated materials 107
translating the headword 36
translation equivalents 36, 67, 70, 114, 150
triglot 15, 18, 34, 53, 59, 64, 71, 91, 187, 199
trilingual 71
trilingual dictionaries 60, 70
trouble merging documents 2, 54
two views of language 100
types of a kind 126

—U—

unaccusative 167
undergoer 21, 127, 152, 166
underline 17
underline bold 50
underline character 50
underline code 51
underline italic 50
underlining affixes 51
underlying forms 22
underlying roots 22
unergative 167
unergative-unaccusative 166
unformatted 25
unifying definition 115
uninitiated user 157
Unit 133
unknown fields 29, 56
unstructured text files 89
unwanted fields 9
UPDATE.CCT 199, 201
UPPER CASE 3
usage 18, 20, 67, 70, 120, 150, 155, 179
usage restrictions 24
user-defined sort orders 7
user-friendly 157

—V—

variant 24, 42, 117, 120, 124, 155, 179
variant forms 23
varieties 142, 144, 146
variety of output options 4
verb class 25
verbal subclasses 166
vernacular 20, 49
vernacular categories 26
vernacular definition 41
vernacular explanations 16
version of WORD 54
visual examples 29
vocabulary 67
vulgar 206
Vwhole 129

—W—

Whole 129
Wierzbicka 40, 115, 157, 162, 164, 170
Windows users 2
WINWORD 1, 54
Wolff and Poedjosoedarmo 154
WORD 1, 3, 9, 53, 54, 58, 61, 63, 64, 95
word class 157
WORD-for-DOS 54
WORD-for-WINDOWS 54
word-level gloss 17, 18, 19, 38
wordlists 72
writing a good definition 39
Wurm and Wilson 24

—Y—

your word processor 3, 54, 55

—Z—

zero-derivation 161, 163
Zgusta 108, 115, 157, 171
zoologists 73, 137, 141
zoology 19